

Towards the Automatic Assessment of Social Experience in in-the-wild Mingling Settings

Vargas Quiros, J.D.

DOI

[10.4233/uuid:6f612795-5d2a-4768-904d-76087d84a0df](https://doi.org/10.4233/uuid:6f612795-5d2a-4768-904d-76087d84a0df)

Publication date

2024

Document Version

Final published version

Citation (APA)

Vargas Quiros, J. D. (2024). *Towards the Automatic Assessment of Social Experience in in-the-wild Mingling Settings*. [Dissertation (TU Delft), Delft University of Technology].
<https://doi.org/10.4233/uuid:6f612795-5d2a-4768-904d-76087d84a0df>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



TOWARDS THE
AUTOMATIC ASSESSMENT
OF SOCIAL EXPERIENCE
IN IN-THE-WILD MINGLING SETTINGS

**TOWARDS THE AUTOMATIC ASSESSMENT OF
SOCIAL EXPERIENCE IN IN-THE-WILD MINGLING
SETTINGS**

TOWARDS THE AUTOMATIC ASSESSMENT OF SOCIAL EXPERIENCE IN IN-THE-WILD MINGLING SETTINGS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Monday, 7 October 2024 at 17:30 o'clock

by

José David VARGAS QUIRÓS

Master of Science in Computing Science, Utrecht University, The Netherlands
Born in Gothenburg, Sweden

This dissertation has been approved by the promotor.

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, promotor.
Dr. H.S. Hung	Delft University of Technology, copromotor.
Dr. L.C. Cabrera-Quiros	Delft University of Technology, copromotor.

Independent members:

Prof. dr. M.A. Neerincx,	Delft University of Technology
Dr. K.P. Truong,	University of Twente
Prof. dr. A. Hanjalic,	Delft University of Technology
Prof. dr. A.A. Salah,	Utrecht University



This work was supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

Printed by: <https://proefschriftmaken.nl>

Cover: *Le Déjeuner des canotiers*. Pierre-Auguste Renoir, 1881.

Style: TU Delft House Style, with modifications by Moritz Beller
<https://github.com/Inventitech/phd-thesis-template>

An electronic version of this dissertation is available at
<http://repository.tudelft.nl/>.

The more I learn, the more I realize how much I don't know.

Albert Einstein

CONTENTS

Summary	xi
Samenvatting	xiii
Acknowledgments	xv
1 Introduction	1
1.1 Social signal processing, and the quest to understand humans in interaction	2
1.1.1 The multiple facets of social signal processing	2
1.2 Social signals as predictors of social experience variables.	3
1.2.1 Social experience variables.	3
1.2.2 Levels of analysis.	5
1.3 The in-the-wild mingling setting	7
1.3.1 Interaction settings and ecological validity	7
1.3.2 Study of in-the-wild mingling	7
1.4 Limitations and challenges in the automatic assessment of social experience	8
1.4.1 Steps of social signal processing research.	9
1.4.2 Data collection: limitations and challenges	9
1.4.3 Limitations in annotation of human behavior	12
1.4.4 Modelling and analysis challenges	14
1.5 Thesis contributions	17
2 Individual and joint body movement as a predictor of attraction in speed dates	21
2.1 Introduction	22
2.2 Attraction and body movement	23
2.2.1 Interpersonal attraction	23
2.2.2 Individual body movement and attraction	24
2.2.3 Synchrony, mimicry, convergence and their role in attraction	24
2.2.4 Measuring synchrony, mimicry and convergence.	26
2.3 Dataset and methods	28
2.3.1 Experiment context	28
2.3.2 Defining the ground truth	29
2.3.3 Feature extraction	31
2.3.4 Dimensionality reduction	35
2.4 Results	36
2.4.1 Body movement and attraction.	36
2.4.2 Joint body movement and attraction	38
2.4.3 Ablation study: feature type importance	40
2.5 Discussion	41

3	No-audio speaking status detection via pose-based filtering and wearable acceleration	45
3.1	Introduction	46
3.2	Related work	48
3.2.1	Visual detection of actions	48
3.2.2	No-audio speech detection	49
3.2.3	Multimodal speech detection	50
3.2.4	Summary	50
3.3	Approach	50
3.3.1	Voice activity detection from video	50
3.3.2	Voice activity estimation from wearable acceleration	54
3.3.3	Multimodal fusion	55
3.4	Datasets	55
3.4.1	Free standing dataset	55
3.4.2	Automatic speaking status annotation	57
3.4.3	Realvad dataset	57
3.4.4	Data pre-processing	58
3.5	Experiments	58
3.5.1	Pose-based filtering	59
3.5.2	Multimodal speaking status detection	62
3.6	Limitations and future work	63
3.7	Conclusions	64
4	ConfLab: concept, dataset, and benchmark for machine analysis of social interactions	65
4.1	Introduction	66
4.2	Related work	67
4.3	Data acquisition	69
4.4	Data annotation	71
4.5	Dataset statistics	73
4.6	Research tasks	74
4.6.1	Person and keypoints detection	74
4.6.2	Speaking status detection	75
4.6.3	F-formation detection	76
4.7	Conclusion and discussion	76
4.A	Hosting, licensing, and organization	79
4.B	Datasheet for ConfLab	81
4.C	Sample participant report	96
4.D	Data capture setup details	97
4.E	Implementation details	97
4.E.1	Person and keypoint detection models	97
4.E.2	F-formation detection	98
4.F	Additional results	98
4.F.1	Person and keypoints detection	98
4.F.2	Speaking status detection	100

4.G	Reproducibility checklist	101
4.G.1	Person and keypoints detection	101
4.G.2	Speaking status detection	101
4.G.3	F-formation detection	102
5	REWIND dataset: speaking status segmentation from body movements in the wild	105
5.1	Introduction	106
5.2	Related work	107
5.2.1	Related datasets and methods in mingling settings	108
5.2.2	Speaking status detection in non-mingling settings	109
5.3	Data acquisition	109
5.3.1	Participant procedure	110
5.3.2	Sensor setup	110
5.3.3	Data collection details	111
5.4	Data annotation	111
5.4.1	Automatic audio-based speaking status Annotation	111
5.4.2	Semi-automatic pose annotations	112
5.5	Dataset statistics	113
5.6	Baselines for automated speaking status segmentation	114
5.6.1	Evaluation setup	114
5.6.2	Video-based speaking status segmentation	115
5.6.3	Pose and acceleration-based speaking status setectors	115
5.6.4	Multimodal speaking status segmentation	115
5.7	Discussion and conclusions	116
5.7.1	Efficacy of pose-based analysis	117
5.A	Network details	118
6	Covfee: extensible web framework for continuous-time annotation	119
6.1	Introduction	120
6.2	Related work	122
6.2.1	Manually annotating keypoints and actions	123
6.2.2	Crowdsourcing annotations	124
6.2.3	Continuous-time annotation	124
6.3	The Covfee framework for continuous annotation	127
6.3.1	The Covfee specification file	128
6.3.2	Online workflow	129
6.3.3	Data privacy and security	131
6.3.4	Crowd-sourcing support	132
6.3.5	Extensibility	133
6.4	Case studies	133
6.4.1	Case study I: keypoint annotation in group interaction settings	133
6.4.2	Case study II: social action annotation	137
6.5	Conclusion and discussion	139
6.A	Covfee specification examples	142
6.B	Covfee web application	143

7	Impact of laughter annotation modality on label quality and model performance	145
7.1	Introduction	145
7.2	Background and related work	148
7.2.1	The study of laughter in interaction	148
7.2.2	Automatic laughter detection, classification, and intensity estimation in the wild.	149
7.2.3	Laughter perception and annotation	149
7.3	Our approach	150
7.4	Dataset.	151
7.5	Methods	152
7.5.1	Laughter candidate generation	152
7.5.2	Annotation of laughter candidates	154
7.5.3	Measuring inter-annotator agreement	157
7.5.4	Automatic, laughter detection, intensity estimation, and segmentation	157
7.6	Results	159
7.6.1	Comparison of human laughter annotation agreement across modalities	159
7.6.2	Effect of labeling modality on supervised laughter tasks	163
7.7	Discussion	165
7.7.1	Limitations.	167
7.A	Annotation experiment details	169
7.A.1	HIT structure	169
7.A.2	Annotation experiment statistics	170
7.B	Segmentation network details	172
8	Discussion and future work	173
8.1	Summary of contributions and findings	173
8.2	Discussion of Implications	176
8.2.1	Data collection	176
8.2.2	Annotation.	178
8.2.3	Modelling and analysis.	179
	Bibliography	183
	Curriculum Vitæ	223
	List of publications	225

SUMMARY

Endowing machines with social competence is not only a science fiction theme. It is also a long-held goal in computer science. Machines have changed how we work, communicate, and do art, science, and engineering, but they have had little effect on one of our core human needs: social interaction. Although digital communication has changed the way we interact with others, machines have arguably done little to enhance the quality of our face-to-face interactions and are seldom seen as tools to help us improve the way we interact with others. This is in part due to their lack of social competence thus far.

A crucial stepping stone towards social competence and the ability to display empathy is the ability to assess social experience. Social experience refers to internal states reflecting an individual's perception of a social situation, like enjoying a conversation or feeling attracted to someone they are interacting with. Social experience variables are hard to study because they are not directly observable and change over time. Researchers must rely on self-reports or third-party assessments (annotations). Algorithms for assessment of social experience generally take one of two approaches: 1) Direct modeling of the relationship between raw/derived signals and experience variables, utilizing sensor readings or outputs of detectors and feature extractors; and 2) intermediate modeling/detection of discrete actions performed during interactions (ie. speaking, laughter, gesturing).

In this thesis dissertation, we focus on in-the-wild mingling setting, where subjects are standing and are free to form and switch conversation groups as they desire. Data collection and annotation are paid special attention due to their relevance in a nascent field and the nuance involved in collecting and annotating social signals. Because the goal is to study machine social perception in real-life settings, interactions are not scripted and instrumentation is kept to a minimum.

We start with work concerned with the direct assessment of social experience, in this case of attraction, by exploring the predictive power of body acceleration. By analyzing accelerometer data from speed dating interactions, we investigate how the intensity and variations in body movement relate to self-reported attraction levels. This study sheds light on the predictive power of synchrony, mimicry, and convergence estimates for predicting attraction, and potentially other constructs related to affiliation.

We then address the detection of speaking, an action of wide interest in social signal processing due to the relevance of turn-taking in social experience. We address the limitations posed by visual cross-contamination in crowded mingling settings. We introduce a model that employs accelerometer readings and body poses to enhance the robustness of speaking status detection in a complex scene, with multiple interactions occurring simultaneously.

The dissertation also presents two novel datasets: ConfLab and REWIND, each serving a unique purpose. ConfLab, collected during a conference, is notable for its annotations of body joints, and improvements to the sensor setup resulting in increased data fidelity. Such methodological contributions to enable efficient and high-quality data collection are increasingly valuable given the scarcity of social interaction datasets, particularly in

mingling settings. REWIND, gathered at a business networking event, stands out with its high-quality individual audio recordings, useful for the cross-modal study of multimodal signals such as speaking or laughter.

In a similar line, we present the Covfee software framework. Covfee challenges existing annotation methodologies by introducing and studying interfaces for continuous annotation for keypoints and actions. This framework was instrumental in efficiently processing the vast amounts of data collected in studies like ConfLab by streamlining the annotation process.

Also building on the Covfee framework, the dissertation culminates in an exploration of laughter annotation across different modalities. By comparing laughter annotations acquired in different conditions, the research highlights the complexities and nuances involved in interpreting social signals across different sensory inputs. We challenge the assumption that laughter intensity should be considered a property of the laughter episode. Instead, we find evidence that laughter evaluations differ significantly depending on the modalities available to the observer and that modalities with higher agreement will not necessarily result in the highest model performance. These results not only contribute to the study of laughter detection but also provide valuable insights for future research on multimodal social signal processing.

In summary, this dissertation weaves together a series of methodological contributions and novel findings, often derived from these new methods, each contributing to further our understanding of how to best train machines for social understanding and competence.

SAMENVATTING

Het toekkenen van machines met sociale vaardigheden en waarnemingsvermogen is niet alleen een sciencefiction thema maar het is ook een langgekoesterd doel in de informatica. Machines hebben de manier waarop we werken en communiceren veranderd, maar ook hoe kunst, wetenschap en techniek bedreven worden. Echter hebben machines weinig effect gehad op een van onze belangrijkste menselijke behoeften: sociale interactie. Hoewel digitale communicatie de manier waarop we met anderen omgaan heeft veranderd, hebben machines weinig gedaan om de kwaliteit van onze persoonlijke interacties te verbeteren. Machines worden ook zelden gezien als hulpmiddelen om ons te helpen de manier waarop we met anderen omgaan te verbeteren. Dit is voor grotendeels te wijten aan hun gebrek aan sociale competentie tot nu toe.

Een cruciale opstap naar sociale competentie en het tonen van empathie is het vermogen om sociale ervaringen te beoordelen. Sociale ervaring verwijst naar interne toestanden die de perceptie van een individu van een sociale situatie weerspiegelen, zoals het genieten van een gesprek of het zich aangetrokken voelen tot iemand waarmee ze omgaan. Sociale ervaringsvariabelen zijn notoir moeilijk te bestuderen omdat ze niet direct waarneembaar zijn en in de loop van de tijd veranderen. Onderzoekers moeten vertrouwen op zelfrapportages of beoordelingen van derden (annotaties). Algoritmen voor de beoordeling van sociale ervaring hanteren over het algemeen een van de volgende twee benaderingen: 1) Directe modellering van de relatie tussen ruwe/afgeleide signalen en ervaringsvariabelen, gebruikmakend van sensormetingen of uitvoer van detectoren en kenmerkextractors; en 2) tussentijdse modellering/detectie van discrete handelingen die worden uitgevoerd tijdens interacties (zoals bijvoorbeeld spreken, lachen, gebaren).

In dit werk richten we ons op een in-the-wild setting, waarin mensen staan en vrij zijn om gespreksgroepen te vormen en te wisselen zoals ze dat willen. Er wordt speciale aandacht besteed aan gegevensverzameling en annotatie vanwege hun relevantie in een opkomend vakgebied en de subtiliteit die gepaard gaat met het verzamelen en annoteren van sociale signalen. Omdat het doel is om automatische sociale perceptie te bestuderen in levensechte omgevingen, zijn interacties niet gescript en wordt gebruik van instrumentatie tot een minimum beperkt.

We beginnen met onderzoek dat zich richt op de directe beoordeling van sociale ervaring, in dit geval aantrekkingskracht, door de voorspellende kracht van lichaamsversnelling te verkennen. Door versnellingsmetergegevens te analyseren van speeddate-interacties onderzoeken we hoe de intensiteit en variaties in lichaamsbewegingen verband houden met zelfgerapporteerde aantrekkingsniveaus. Deze studie werpt licht op de voorspellende kracht van synchronie, mimicry en convergentieschattingen voor het voorspellen van aantrekkingskracht, en mogelijk andere constructen die verband houden met verbondenheid.

Vervolgens richten we ons op het detecteren van spreken, een handeling die breed interesse wekt in de verwerking van sociale signalen vanwege de relevantie van beurtwisselingen in sociale ervaring. We behandelen de beperkingen die worden veroorzaakt

door visuele kruisbesmetting in drukke informele omgevingen. Er wordt een model geïntroduceerd dat gebruik maakt van versnellingsmetergegevens en lichaamshoudingen om de robuustheid van de detectie van spreekstatus te verbeteren in een complexe omgeving waar meerdere interacties tegelijkertijd plaatsvinden.

Het proefschrift presenteert ook twee nieuwe datasets: Conflab en REWIND, elk met een uniek doel. Conflab, verzameld tijdens een conferentie, is opmerkelijk vanwege de annotaties van lichaamsgewrichten en verbeteringen aan de sensoropstelling die leiden tot een verhoogde betrouwbaarheid van de gegevens. Dergelijke methodologische bijdragen om efficiënte en kwalitatief hoogwaardige gegevensverzameling mogelijk te maken, worden steeds waardevoller gezien de schaarste aan datasets over sociale interactie, vooral in informele settings. REWIND, verzameld tijdens een zakelijk netwerkevenement, onderscheidt zich door zijn hoogwaardige kwaliteit van individuele audio-opnames, bijzonder nuttig zijn voor de cross-modale studie van multimodale signalen zoals spreken of lachen.

In dezelfde lijn presenteert het proefschrift het Covfee software framework. Covfee daagt bestaande annotatiemethodologieën uit door interfaces voor continue annotatie van sleutelpunten en acties te introduceren en te bestuderen. Dit framework speelde een cruciale rol bij het efficiënt verwerken van de enorme hoeveelheden gegevens verzameld in studies zoals Conflab door het annotatieproces te streamlijnen.

Ook voortbouwend op het Covfee-framework, bereikt het proefschrift een hoogtepunt in het onderzoeken van annotatie van lachen over verschillende modaliteiten. Door lachannotaties te vergelijken die zijn verkregen onder verschillende omstandigheden, belicht het onderzoek de complexiteit en nuances die gepaard gaan met het interpreteren van sociale signalen via verschillende sensorische input. Wij betwisten de aanname dat de intensiteit van het lachen moet worden beschouwd als een eigenschap van de lachepisode. In plaats daarvan vinden we bewijs dat lachbeoordelingen aanzienlijk verschillen, afhankelijk van de modaliteiten die beschikbaar zijn voor de waarnemer, en dat de modaliteiten met een hogere overeenstemming niet noodzakelijkerwijs leiden tot de hoogste modelprestaties. Deze resultaten dragen niet alleen bij aan het gebied van lachdetectie, maar bieden ook waardevolle inzichten voor toekomstig onderzoek naar multimodale sociale signaalverwerking.

Samenvattend verbind dit proefschrift een reeks methodologische bijdragen en nieuwe bevindingen, vaak afgeleid van deze nieuwe methoden, die elk bijdragen aan een verdere ontwikkeling van ons begrip van hoe we machines het beste kunnen trainen voor sociaal begrip en competentie.

ACKNOWLEDGMENTS

My PhD is finished, and many special people deserve a warm thank you for making it possible.

First of all, I want to thank my advisors and promotor. Thank you Hayley for giving me this opportunity in the first place and then your unwavering support and optimism throughout. Thank you for constantly challenging my often outrageous ideas for four straight years. Thanks to this I now carry a little Hayley in my head who knows how to question most of my ideas. It might sound like trauma but it's just critical thinking. Thank you Marcel for your wise and practical advice when it was most needed. Thank you Laura for eagerly sharing your experience and enthusiasm.

Thank you SPC lab mates: Stephanie for always being willing to listen and deliver honest advice and ideas. Chirag for tirelessly sharing your wisdom; which both made me wiser and led me to discover ANC technology (Active Noise Cancelling). Ekin for your help especially when I started and for showing me the limits of ANC technology (Active Noise Cancelling). Laura for all your help and selflessness when I began the PhD and you were about to finish and for always offering your listening ear. Tiffany for your enthusiasm and unique points of view. Bernd for your insights and for being a solid German friend, with all the advantages and disadvantages thereof.

Special thanks to my friends in Delft who helped make this journey fun: Masha for your support and out-of-the-box ideas, Jay for supporting and amplifying Masha's out-of-the-box ideas; Dmitry for making me feel sane with your camera obsession; Anurag for supporting my very successful photography and tennis careers; and Bernd for being a solid German friend, with all the advantages and disadvantages thereof.

I extend the acknowledgement to all former colleagues at the Pattern Recognition and Bioinformatics Lab; particularly to Yeshwanth for being the nicest; to Ziqi for causing chaos when needed; to Tom for broadcasting the chaos; to Arman for always having something interesting to talk about; to Osman and Yancong for your solid career and fitness advice. To Ruud and Bart for their technical support and for letting me have a second monitor.

A special thanks to the students I had the pleasure to work with: Öykü Kapcak, Ailin Liu, and the many students I supervised in bachelor's and master's courses. I learnt a lot from you. I hope it was mutual. To Martha and Odette for your commitment to teaching multimedia analysis and for making TAing a rewarding experience.

The greatest acknowledgement is for my parents William and Shirley and brother Carlos. Thank you for your unconditional and irreplaceable love and support. Many thanks to my numerous aunts and grandma for always cheering for me from afar. To my life-long friend Emanuel and the many friends from back in Costa Rica who I unfortunately seldom see but who I know would be happy to read this.

*Jose
Delft, September 2024*

1

INTRODUCTION

Social perception includes the ability to assess experience from social signals. Research into developing machine social perception involves data collection, annotation, modeling, and analysis stages. Social signals are generally captured via video cameras, microphones, and wearable sensors. Social experience measures are normally annotated via self-reports or third-party questionnaires. Regarding modeling, two main approaches can be distinguished. One involves the detection of social actions such as speaking or laughing as an intermediate step. A second approach is to directly model the outcome variables from social signals. Finally, the analysis stage may take various forms but often focuses either on validating models or in studying social signals with set models. Each of the stages, from data collection to analysis faces important challenges or opportunities. In this chapter, we 1) dive into the conceptualization and contextualization of the problem of automatic assessment of social experience; 2) review and discuss many of the open challenges faced in data collection, annotation, modeling, and analysis 3) conceptualize each of the chapters of this thesis with respect to these challenges.

1.1 SOCIAL SIGNAL PROCESSING, AND THE QUEST TO UNDERSTAND HUMANS IN INTERACTION

Social signals are everywhere in our social lives: from a baby crying to be fed, to a teenager flirting to signal attraction, or a grandmother hugging her grandson to display affection. Some of the first definitions equated social signals to *actions* occurring in a social context, with the additional property that they influence the behavior or internal state of others (ie. they are signals). Social signals, however, are not restricted to atomic, named actions such as crying or hugging. Factors like physical appearance (ie. attractiveness, or signs of old age), body pose, and interpersonal distance also communicate social information in an interaction setting. Definitions inclusive of this breadth include the one by Poggi and D’Errico, who define a *signal* as “any perceivable stimulus from which a system can draw some meaning” [1, p. 189], and a *social signal* as “a communicative or informative signal which, either directly or indirectly, provides information about social facts, that is, about social interactions, social attitudes, social relations and social emotions” [1].

Burgoon et al. [2] identify some consensus characteristics of social signals: a) they are observable, b) they produce changes in others, and c) these changes are not random, but follow laws and principles. They define social signal processing as a computing domain aimed at modeling, analysis and synthesis of social signals in human-human and human-machine interaction [2]. We will now dive into the details of what *social signal processing* means in practice.

1.1.1 THE MULTIPLE FACETS OF SOCIAL SIGNAL PROCESSING

It is important to note that the definition above for *social signal processing* can be understood in two ways: a) as a computational methodology (models and analysis techniques), which can be used to study social signals, and b) as a scientific field, whose object of study is the computational modeling, analysis, and synthesis of social signals.

Regarding the first, social signals have been studied by many disciplines. Only in the case of laughter, for example, psychological research studies theories and models related to its causes, form, functions, and contexts [3–8]. Linguists have been concerned with the functions and meaning of laughter in dialogue [9–13]. Biology has concerned itself with how laughter originated and with the issue of *nature versus nurture* [14–16]. Meanwhile, neuroscientists are concerned with where and how laughter originates and is perceived in the brain [17, 18], while medicine has studied the health effects of laughter in the body [19]. Any of these disciplines concerned with the study of social signals can, and often do make use of computational modeling and analysis of social signals (ie. social signal processing) to reach its conclusions.

As a scientific field, social signal processing, widely considered part of computer science is often concerned with aiding these disciplines or with addressing research questions *belonging* to them using its own methods. The application of machine learning enables computer scientists to study research questions not traditionally considered by other disciplines concerned with social signals. The behavioral psychologist, for example, rarely models behavior from raw data directly and instead makes use of a human behavioral coding step to study discrete actions [20]. This allows the psychologist to study the link between the occurrence of a behavior (eg. laughter) and another variable of interest (eg.

the enjoyment of an interaction). In contrast, the social signal processing practitioner will often be concerned with modeling both variables directly from recordings of social signals like video, audio, or wearable signals. Psychological research may be used to inspire and determine their research questions, or to interpret their findings.

The social signal processing field is often understood to have its own goal: aiding human experience through technology by endowing machines with social intelligence [21]. This is most evident in the synthesis task, which has as its end goal the development of artificial agents capable of imitating human behavior [22]. Research concerned with modeling and analysis is also commonly motivated by the ultimate desire to equip technological applications with the ability to *understand* human social signals, and act upon them [23, 24]. Technological applications may include therapy and care support [25, 26], social recommender / support systems [27], or artificial agents with a variety of goals [28–34]. In other words, social signal processing has the goal of studying technological interventions.

In line with the two interpretations above, here we use the term *social signal processing* to refer to the modeling, analysis, and synthesis of social signals, regardless of discipline. Challenges towards these goals are shared by multiple disciplines, and ultimately research questions, not discipline boundaries, should drive the methods used in any one study. Given this landscape, contributions to social signal processing can be roughly classified into two camps:

Methodological contributions , where we seek to improve the methods used in the study of social signals. This includes but is not limited to: data collection methods, annotation methods, machine learning methods, and statistical analysis methods.

Social signal facts , where we discover scientific facts related to social signals, possibly linking them to other variables. Research questions in this camp can often be considered to *belong* to other disciplines such as behavioral psychology, due to its object of study: human thought and behavior.

Although not a rule, it could be said that the *field of social signal processing* is the one primarily concerned with methodological contributions. Social signal facts, meanwhile, are pursued both by computer scientists and practitioners of other disciplines, but social signal processing is the field best positioned to make use of machine learning for the task.

1.2 SOCIAL SIGNALS AS PREDICTORS OF SOCIAL EXPERIENCE VARIABLES

While social signals can be studied in isolation, they are most often studied in relation to another variable. Demographic variables like personality [35–37], age [38], gender [38], relationship status [39] or culture [40–42] may come to mind. In particular, a significant part of social signal processing literature has focused on the link between social signals and social experience variables.

1.2.1 SOCIAL EXPERIENCE VARIABLES

We use the term *social experience variable* loosely to refer to the internal states of the individual related to their assessment of a social situation. Intuitively, assessing social experience refers to answering questions such as:

- is the subject having a good time?
- does the subject like who s/he is talking to?
- are the subjects attracted to each other?

Such target variables align with social signal processing's vision of *aiding humans in interaction*, since the goal of technological interventions may be framed as the optimization of one of these variables.

Being internal states, social experience variables are not directly observable by the researcher. Therefore, in practice, social experience is always studied indirectly, through:

Self-reports Subjects provide reports of their experience after an event, usually by filling in a questionnaire. Despite coming directly from the subject, self-reports cannot be considered objective. Appraisal theory establishes that the meaning of an event for an individual is not constant in time [43], as re-appraisals occur when an event is recalled. The act of providing a report itself constitutes the triggering of an appraisal process. In other words, social experience is not a constant to be extracted from the individual, but a changing variable that cannot be measured without side effects. While these nuances are important to understand, self-reports are widely used in practice to quantify social experience, as they are the most straightforward way to obtain experience information directly from subjects. However, self-reporting is generally limited to summary evaluations of a conversation or interaction (ie. low temporal resolution) for practical reasons.

Third-party annotations Here the target variable is assessed by an observer, who, directly or indirectly answers a question such as: *does the subject appear to be having a good time?* This is the prevalent approach taken by the affective computing community [44], concerned with the modeling, analysis, and synthesis of *emotional expressions*. While the correspondence between what is displayed in the body (ie. affective expressions) and internal states has been a heated topic of debate, it is clear that third-party annotations are far from direct measurements of internal states. In addition to the possible dissociation between expression and thought, the measured variable is affected by the perceiver's biases and their context. Furthermore, annotations are done on recordings, adding a layer between behavior and observer. Rather than directly assessing internal states (ie. thoughts or feelings), third-party annotations are useful for endowing machines with the ability to assess expressions and behaviors as humans do. Third-party annotations have the advantage (over self-reporting) that they can be obtained from media at any moment, and at higher temporal resolutions.

Despite these nuances, we will use the term *social experience variables* to refer to both of the above. Social experience variables of interest in social signal processing have included enjoyment [45, 46], engagement [24, 37, 47–50], involvement [51–54], attraction [55–57] or affective dimensions [44].

1.2.2 LEVELS OF ANALYSIS

An important dimension, helpful to situate work in social signal processing is that of *level of analysis*, used here to refer loosely to the level of abstraction of the variables considered. At the lowest level, we have raw data like video, audio, or wearable accelerometer readings. At a *higher level* there are social experience variables like enjoyment or engagement. At an intermediate level, we might have social actions, which can be annotated or automatically detected from raw data, and used to infer social experience variables. In practice, we are often interested in establishing relationships across these levels: *can we detect laughter from raw data? can laughter occurrences be used to assess enjoyment in an interaction*. Skipping levels is also possible: *can we detect enjoyment directly from raw data, without first detecting actions?*

Figure 1.1 is a diagram of these levels. In practice, most social signal modeling and analysis tends to be *bottom-up*, by training a model to infer actions or social experience from raw data and/or derived signals. Some synthesis work can be understood to be *top-down*, where the goal is to generate low-level social signals with certain characteristics. There are no hard rules, however, and contributions within a single level are also possible.

Figure 1.1 highlights the two main approaches to predicting social experience. In the first, we model the relationship between raw / derived (low-level) signals and experience variables directly [44, 49, 55, 56, 58]. Input signals are usually raw sensor readings (video, audio, accelerometer), or the output of detectors (eg. poses, facial keypoints) or feature extractors. Psychological research provided a reason to believe that it is possible to learn about social experience from raw social signal readings. This is not only because they hold information about *social actions* like laughter. Inter-personal phenomena like mimicry (the tendency to copy behavior from the interaction partner), have been linked to more favorable evaluations from an interaction partner [59], attraction [60, 61], higher ratings of smoothness of the interaction [62] and an increased desire to be liked by an interaction partner [59, 63]. Different communities have addressed the challenge of designing features and models capable of capturing these phenomena [64–74].

A second approach, more in line with research in behavioral psychology, is to start by modeling or acquiring annotations of discrete actions that humans perform in an interaction, to model experience variables as a function of those actions [75]. In addition to potentially better performance, using actions as an intermediate step may provide insights about the physical manifestations of social experience that are not straightforward to reach with models operating directly on raw data. The use of actions as an intermediate step requires assumptions about what kinds of actions are relevant for modeling a particular variable (eg. is yawning relevant to predicting if somebody is enjoying a conversation?). Although it cannot be assumed that actions are useful in general, certain common social actions like speaking, patterns in speaking activity (turn-taking), and laughter have been found to be linked to a wide range of variables, including enjoyment, engagement, and attraction [45, 75–78].

Within the social signal processing field, the direct modeling of a target variable is a popular approach. Event-based models that incorporate actions, though far less popular, have received some attention [45, 79]. In particular, models have been developed to address speaking status detection [80, 81], laughter [82–84], back-channeling, and other social actions [85] from multiple modalities. Although they are currently seldom used as part

Figure 1.1: Levels of analysis in the study of social experience from social signals. Most work in social signal processing will seek to establish relationships within or between these levels.



of a larger system, the development of action detectors holds the promise of potentially replacing manual behavioral coding/annotation in future research on social actions.

1.3 THE IN-THE-WILD MINGLING SETTING

In section 1.1.1 we talked about social signal processing contributions as being either related to the study of social phenomena (social signal facts) or methodological. In Section 1.2 we narrowed the scope to the problem of assessing social experience variables from social signals. A third distinction of importance in social signal processing work is the *setting* where social signals are collected. In this section, we introduce the importance of social interaction setting and further narrow the scope to the main setting of interest in this work: the *in-the-wild mingling* setting.

1.3.1 INTERACTION SETTINGS AND ECOLOGICAL VALIDITY

We use the term *setting* loosely to describe more than just the location or kind of event where the data was collected. The setting includes other information of relevance when analyzing social signals such as: a) was the interaction in dyads, fixed-size groups, or larger groups? b) did groups interact separately or all in the same space? c) were participants able to freely switch groups? d) were subjects sitting standing or otherwise? e) were interactions scripted or naturally occurring? f) did subjects know each other beforehand?

Answers to these questions may involve a significant amount of nuance for any particular dataset, but they are relevant because they determine the conclusions that can be drawn from an analysis. Social signal processing work can be found on a variety of settings: from dyads interacting in a lab, surrounded by instruments, and following a script; to freely-interacting crowds in a real event, only recorded by overhead cameras. These two examples highlight an important dimension of social interaction setting: their *ecological validity* [86]. This term has been used in psychology to refer to the expectation that “findings in the laboratory will be able to generalize to the real world outside the laboratory” [87, p. 466]. What makes an experiment *ecologically valid* is therefore a function of its research questions. In *in-the-wild* settings, where data is not collected in a laboratory, the term refers to the degree to which the conditions in which the data is collected resemble conditions occurring in real-life interaction.

The use of a script, constraints on group sizes, and instrumentation are all factors that may affect the ecological validity of a study. Studies are said to have high ecological validity when these factors are minimized (conditions are kept as *naturalistic* as possible). Some research questions, however, require the use of less ecologically valid settings if, for example, the interaction must be steered towards a particular goal, or the raw data of interest requires invasive instrumentation. Because most real-life interactions are not recorded for research purposes, the fact that subjects know being recorded alone may affect the ecological validity of a study. This is, however, a necessary evil, as law and ethics establish that data subjects should be aware when their personal data are recorded.

1.3.2 STUDY OF IN-THE-WILD MINGLING

A line of work in the social signal processing community (including most of the work in this thesis) has focused on the *in-the-wild mingling setting*, also sometimes referred to as *cocktail*

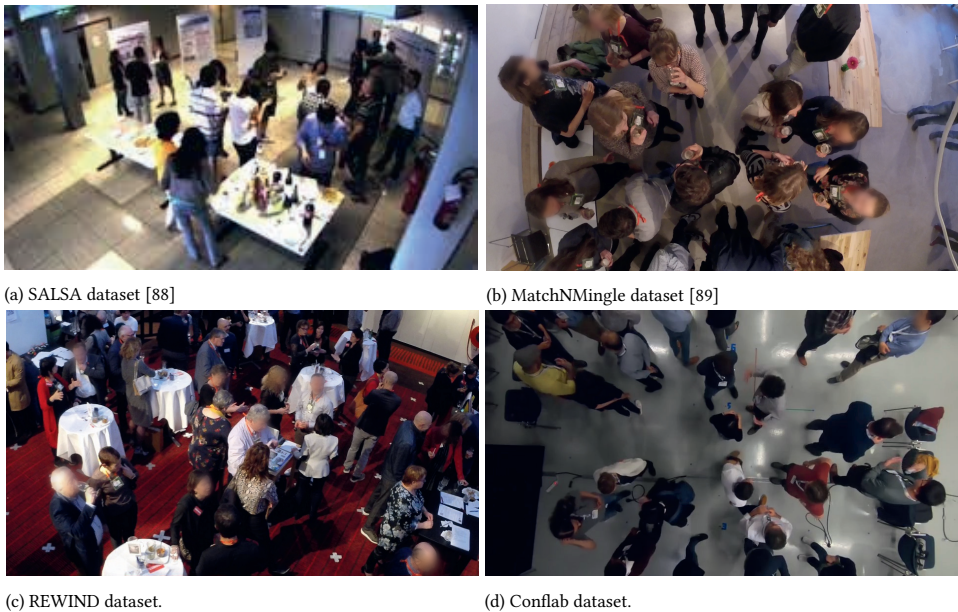


Figure 1.2: Examples of in-the-wild mingling datasets, where a small crowd interacts freely.

party setting [88, 89]. In this setting, subjects are usually standing and are free to *mingle*, that is, to form and switch conversation groups as they desire (Figure 1.2). Since subjects are given no objective, they are free to follow their own goals and desires when choosing conversation partners. The formation and evolution of *F-formations*, spatial arrangements roughly (but not strictly) corresponding to conversation groups, is characteristic of this type of setting, and has been one of the main subjects of study in this setting [90–94].

The social signal processing community has further prioritized the ecological validity in research of mingling settings by collecting datasets in real-life events (not planned only for data collection) [88–90]. Non-invasive instrumentation has been used; generally video and optionally wearable devices for recording of body acceleration and proximity information. Individual audio is most commonly not recorded, due to the technical and privacy challenges of recording audio for a small crowd [95].

1.4 LIMITATIONS AND CHALLENGES IN THE AUTOMATIC ASSESSMENT OF SOCIAL EXPERIENCE IN THE WILD

In the previous sections, we narrowed the scope to a particular task (prediction of social experience variables from social signals) and setting (in-the-wild mingling scenarios). Now, we focus on the challenges towards assessing social experience variables in mingling settings. Some challenges are general to the assessment of social experience (across settings), others are general to the analysis of mingling settings (across tasks), and others are specific to the assessment of social experience in mingling settings.

1.4.1 STEPS OF SOCIAL SIGNAL PROCESSING RESEARCH

Researching how machines answer questions about social experience involves solving multiple challenges in addition to the modeling; starting with the technical, logistic, and privacy challenges of acquiring the necessary data. Regarding its process, most work within social signal processing follows the following steps:

- 1. Data collection** Recording a social scene and (optionally) obtaining self-reports.
- 2. Annotation** This may be a combination of manual and automatic processes. Normally it involves detecting people in the scene. Often it also includes annotating their actions, affective displays, or perceived social experience, depending on the goals of the study. Self-reports obtained in (1) may act as a substitute for annotations but are generally of much lower time resolution.
- 3. Modeling** Normally, we model a target signal or experience variable as a function of multiple input signals, usually lower-level social signals (eg. pose information, speech prosody). Target signals may range from individual, concrete actions like *laughter*, to more abstract group constructs like *group enjoyment*. However, the target may also be directly recorded by a sensor (eg. body pose, speech intensity). There need not be distinct targets and inputs, however, since we might be interested in unsupervised problems.
- 4. Analysis** We perform an analysis using our model. This might range from performance comparisons to validate model improvements (methodological contributions), to analyses seeking an understanding of the underlying phenomena (social signal facts) or the way they are captured in a model.

In the following sections, we address some important limitations and challenges in the current practice of social signal processing, for each step of this process.

1.4.2 DATA COLLECTION: LIMITATIONS AND CHALLENGES

Collecting data on human behavior is a major undertaking, especially when the goal is to collect enough data and of sufficient quality, to answer novel research questions in an ecologically valid setting. Improving along these two dimensions: the quality and quantity, or scale of the data are the two major challenges faced by data collection practices.

DATA FIDELITY LIMITATIONS AND CHALLENGES

The use of non-invasive sensing modalities involves a trade-off, in the form of less direct access to social signals. In-lab data collection with a few subjects at a time can often use high-fidelity setups, with high-quality audio, cameras aimed at each subject, and multiple wearable sensors. In ecologically-valid mingling settings, several simplifications have become standard practice. Instead of having a camera focused on each subject's face, mingling settings are recorded via cameras focusing on large parts of the scene. Furthermore, individual audio is often not recorded, due to the privacy and logistic challenges involved in recording audio for a small crowd. Lack of access to audio largely excludes the category of verbal behavior from being annotated and studied in mingling settings. Opportunities

for the study of verbal behavior and the relationship between verbal and non-verbal behavior are lost. Recently, chest-worn wearable sensors have been used to record individual acceleration and proximity information. Acceleration signals are useful for the detection of speaking, with performance superior to that of video. However, proximity information has not been fully leveraged for the detection of social groups.

This creates challenges in at least three directions:

Person detection and tracking in videos Overhead and side-elevated views have been used in existing mingling datasets [88, 89, 95]. This has created the challenge of detecting and tracking people in the scene [96]. Person detection and pose estimation from frontal views have enjoyed significant success thanks to deep learning models trained on large image datasets [97–99]. Frontal view pose detectors have been used to obtain poses used as inputs for modeling and analysis [100]. However, models trained on frontal viewpoints do not transfer well to overhead views [95]. The smaller size of mingling datasets makes retraining infeasible. Solving this problem will likely require a combination of technical solutions (eg. transfer learning to take advantage of frontal view data) and the annotation of more data of overhead views.

Audio sensing for mingling settings . Wireless microphones, though high-quality, are hard to scale past a few dozen due to bandwidth constraints. Furthermore, introducing individual microphones adds an invasive sensor that has to be worn by each participant in the interaction. Wearable devices are a promising development in this area. Through custom designs using internal memory and low-power microcontrollers, these devices can store audio for hours of interaction. To offset privacy concerns, decimated audio, still useful for the detection of speaking activity, can be stored instead [101, 102]. Dealing with the significant amount of background (cocktail party) noise in mingling settings is still an open challenge.

Annotation of social actions from body movement Although the unavailability of audio excludes the direct annotation of verbal behavior, not all is lost when it comes to verbal behavior since actions such as speech or laughter have manifestations in non-verbal behavior that have been annotated from body movement as captured in video [89, 95, 103]. Although body-movement-based annotations are not a substitute for audio-based ones in terms of their scope, the possibility of annotating (and automatically detecting) certain behaviors like speaking or laughter from body movements alone opens the door for a range of research questions. Both speaking status and laughter, for example, have been linked to the social experience variables *enjoyment* and *engagement* [45, 75, 76]. It is unclear, however, whether the process of annotating verbal behavior from non-verbal information is accurate enough to justify its use.

Wearable sensing and wearable data processing There are a variety of challenges to address regarding wearable sensing, such as: finding ways to leverage proximity information for detection of groups. Improving the time and measurement resolution of wearable sensors is also likely to improve the detection of subtle behaviors. The ergonomics and battery life of wearable sensors have also been identified as a point of improvement [104].

PRIVACY, AND THE CHALLENGE OF SCALE

Most currently available mingling datasets correspond to the recording of a single event, with specific settings and demographics of data subjects. Furthermore, datasets in social signal processing are multi-purpose, meaning that the same dataset is used to answer various research questions, usually involving training models for multiple tasks. The study of social signals in mingling events is therefore limited to a handful of mingling events, with a few dozen participants each [95]. This makes it impossible to test findings in a wide variety of mingling contexts. Furthermore, the sub-field is currently not in a position to generalize its findings to human (sub-)populations due to a combination of large standard errors and the fact that current datasets are not designed to capture an unbiased sample of a population. This affects its ability to contribute to the development of psychological theory (or that of other fields). For the same reasons, the difficulty in collecting new datasets is a limitation to the study of technological interventions, which would normally require the study of multiple population samples (under different conditions); large enough to reach statistical significance on a potential effect of interest.

The terms *scale* or *size* of a dataset can be interpreted in multiple ways. We can refer to the number of events recorded, the duration of the events, or the number of subjects in an event, among others. The answer to precisely what *scale* is needed in a dataset is research-question-specific. However, given the multi-purpose quality of social signal processing datasets, data collection would benefit from improvements along all mentioned dimensions of *scale*. The issue of scale is not unique to mingling settings but can be extended to most settings studied in the social signal processing community. However, the mingling setting faces some unique challenges when it comes to scaling the size of its datasets. Privacy laws, like the well-known General Data Protection Regulation (GDPR) of the European Union, play a central role in shaping these challenges and the resulting opportunities, and we address them next.

GDPR defines *personal data* as “any information relating to an identified or identifiable natural person” [105]. The use of the term *identifiable* means that the data does not need to explicitly contain the name or details of the data subject to be considered personal. Face shots, videos, and audio recordings, which can be used to identify individuals also fall under the definition of personal data. Only personal data that “has been rendered anonymous in such a manner that the individual is no longer identifiable” is no longer considered personal [105]. Crucially, most collection of personal data for research purposes requires *informed consent* by the individual. Meanwhile, non-personal data is not protected by GDPR and can, in principle be freely recorded and shared (unless in conflict with other laws).

The landscape created by GDPR has created challenges for the scaling of the data collection process along at least two axes:

Challenges to use of video and audio The GDPR definitions presented above imply that, in mingling events where the interaction scene is recorded by cameras, all of the subjects within the recorded area must provide their informed consent. When recording a real-life event, the opposition of a single participant to be recorded may pose a challenge, as event organizers may be unwilling to put even one of their participants in a difficult position.

The recording of audio is similarly nuanced. While personal microphones would ideally record only the wearer, in practice, in a mingling setting they capture the

voices (and information relayed through the wearer) from surrounding speakers. Opportunities exist to reduce cross-contamination from speakers other than the wearer. This challenge must be addressed through hardware, at the recording stage, since post-hoc source separation still constitutes the processing of personal data from all subjects recorded. However, the leakage of personal information from subjects not part of the study through the words of the wearer is a real possibility.

Challenges with alternative modalities Not all modalities are created equal when it comes to privacy. While the use of cameras requires consent from all participants in the frame of the cameras, some wearable signals like acceleration record only the wearer, and may be used only by some of the participants in the scene. Such modalities are more suitable for scaling up the number of participants and widening the range and number of events in which data is collected. This offers enormous flexibility by avoiding the need for *unanimous consent* by data subjects. This comes at a cost, as certain types of analysis (eg. group-level analysis) are impaired by partial data. Furthermore, excluding video and audio modalities from data collection runs into the issue of how to annotate behavior. Current video-centric annotation methodologies do not have an effective answer to this question, making video recordings essential in mingling data collection. The challenge of video-less annotation may potentially be addressed through a combination of audio and limited pose information captured by wearable sensors. Perception and annotation of social behavior from pose information has already received some attention within the community, although not motivated from this angle [106–108].

1.4.3 LIMITATIONS IN ANNOTATION OF HUMAN BEHAVIOR

Of the steps involved in social signal processing research (Section 1.4.1), the annotation of human behavior datasets is one of the most demanding of the researcher’s time. Even when annotations are not performed directly by the researcher, the hurdles involved in finding and deploying the right annotation tools, finding and organizing suitable annotators, and conducting and verifying the annotation process make annotation a big undertaking. In the previous section, we addressed challenges faced at the data collection stage. In this section, we address the challenges faced during annotation, which currently usually revolves around the video modality.

THE CHALLENGE OF SCALE

Annotation of human behavior faces the same challenge as data collection: how to support the creation of datasets at a larger scale, to cover a larger diversity of subjects, demographics, and data collection settings. This is equivalent to the challenge of making data annotation more time-efficient without negatively affecting annotation quality. Some of the concrete challenges faced are the following:

Subject localization fidelity and time-efficiency Limitations in annotation of mingling datasets start with the issue of (spatially and temporally) localizing subjects in videos. As mentioned in Section 1.4.2, the ideal is to perform this automatically via person detectors and pose estimators. Due to the under-performance of overhead-view detectors and estimators compared to frontal view ones studies currently deal

with the issue in other ways. Manual pose annotation had, until recently, not been attempted in mingling datasets [95], due to the amount of labor required using existing techniques. Instead, researchers have resorted to bounding boxes as the standard way to localize subjects in mingling datasets. It is clear, however, that bounding boxes are not as information-rich as pose detections, which can be used as model inputs in isolation, to develop joint pose-video action detectors, or to obtain bounding boxes.

Time-efficiency of action annotation The challenges continue when annotating actions or social experience variables. Here, the annotation process might take a different form depending on the type of the target variable. Actions are generally represented as binary time series; either occurring or not at any point in time. Current annotation techniques consist of the user drawing an interval on top of a timeline, or setting a binary flag on and off to indicate the time interval of the action. Such approaches to annotation, however, are most appropriate for the annotation of sparsely occurring actions in comparatively short videos (common in computer vision, where they originated). The process becomes slow and tedious when annotating common actions such as speaking or gesturing behavior for hour-long interactions, common in social signal processing datasets [88, 109, 110]. Furthermore, it is hard to estimate the time necessary for annotation, since the input time may vary widely per annotator.

In contrast to actions, variables like enjoyment, engagement, or affective dimensions are normally modeled as continuous variables. The issue of their annotation has been primarily addressed by the affective computing community [111–113]. Dimensions of affect are commonly annotated via continuous-time annotation, a process where the media is annotated while being watched by the annotator, usually without pauses. To this end, the annotator controls a level indicator (with their keyboard, mouse, touchscreen or even gamepad) to, for example, rate the level of arousal of a target subject [114–116]. This means that annotations are usually done in real-time, giving continuous annotation the advantage of being predictable in terms of human labor time. Continuous annotation does have the notable drawback of being affected by annotator reaction delay, a topic which has received attention in the affective computing community [117–119].

Addressing the limitations of current techniques requires research of the annotation processes themselves. Adapting continuous-time annotation techniques to actions is particularly promising since it addresses both limitations mentioned. However, the suitability of these techniques is unverified and the issue of annotator delay must be addressed for discrete actions.

Lack of modern, web-based annotation platforms Finally, annotation of social interaction datasets requires software. Given the relatively niche quality of social interaction datasets in comparison to datasets used in larger fields such as computer vision, software addressing this issue has experienced a significantly lower degree of development. This includes software designed for the annotation of long-duration, non-sparse actions (as explained above), and software for the continuous annotation of social experience variables. Although software exists that is geared towards the annotation of human behavior [120, 121], there are two important limitations to

existing software. The first regards the lack of web and crowd-sourcing support. This affects the ability to scale annotation processes through access to a large worker pool. For example, the first continuous annotation tool with web (browser) support became available only recently [116]. A second limitation regards the flexibility of current software, which implements specific annotation interfaces but does not provide a framework for efficient implementation of new techniques or interfaces. Therefore, experimenting with new annotation tasks and techniques requires the researcher to start from scratch, and solve many engineering challenges in the process [110]. Development of such flexible platforms, however, requires an even larger initial engineering effort. Design considerations are vast, as both the researcher and annotator are users of the software and should be satisfied.

THE ISSUE OF VALIDITY

A second latent challenge in the current study of social experience in interaction is the validation and understanding of current annotation practices. The study of video-based annotation has already been mentioned in Section 1.4.2. Although video-based annotation techniques have been used to annotate mingling datasets [89], the process itself has never been validated or studied. In general, the question of how modality and data recording quality (which affect access to behavior) affect annotation in terms of validity, inter-annotator agreement, and model performance is an open one.

A second issue regarding validity concerns third-party annotation (Section 1.2.1). Most social experience annotations make use of third-party, or observer ratings. However, the fact that the latent goal of social signal processing is to ultimately improve human experience raises the question of how much observer ratings can approximate self-ratings. Despite the widespread use of observer ratings, this question remains largely unexplored. This problem is ripe with nuance, as appraisal theory would predict that the meaning of an event for an individual should not be expected to be constant in time. The study of memory and appraisal of social events is therefore likely to play a key role in providing an answer to this question.

1.4.4 MODELLING AND ANALYSIS CHALLENGES

As explained in Section 1.2, approaches towards assessing human experience in interaction can be classified into two camps: the direct assessment from raw data, and assessment through the intermediate step of detecting social actions (eg. laughter). Each of these approaches faces its own set of challenges. Here, we elaborate on some of these challenges and the opportunities they represent.

SUBTLETY, OCCLUSION, AND CROSS-CONTAMINATION

As with annotation (Section 1.4.3), the detection of social actions involves challenges that are specific to human behavior data. Following, we elaborate on three of these challenges:

The subtlety of social actions One challenge in action recognition of social actions regards the subtle nature of most social actions. While datasets for action recognition in computer vision traditionally contain clear and distinct actions such as sports activities, social signal processing is interested in detecting subtle social cues. Actions as subtle as eyebrow frowns convey social information in interaction [122, 123].

Even laughter, misunderstood popularly to be always loud and clear, occurs most commonly with low intensity in conversations [22]. This means that methods that work well for generic action recognition tasks do not necessarily work well for detecting social actions, and are rarely tested on social action recognition tasks. At the same time, methods designed specifically for the detection of social actions are generally evaluated on fewer datasets and actions, often a single one.

Occlusion and cross-contamination in videos In-the-wild mingling datasets, where videos are often recorded via overhead or side-elevated views, additionally present the challenge of significant occlusion of the target subjects. Occlusion may take different forms. Subjects' faces, often informative of social actions, may be occluded by their own body (self-occlusion) when the subjects face away from the camera. Occlusion by other subjects is also possible. This second type of occlusion can also be understood as *cross-contamination*: the contamination of one subject's signals with another subject's signals. This can hurt action detectors, especially when detecting social actions that can be expected to be inversely correlated, such as speaking status (where most often one subject speaks while the other(s) listen).

Wearable modalities that record only (or mainly) the wearer such as acceleration, in addition to their privacy-related advantages (Section 1.4.2) also have the advantage of not being affected by occlusion and cross-contamination. Using these modalities in isolation or together with video is an opportunity for creating models robust to these factors.

The problem can also be addressed by enhancing the robustness of video methods to cross-contamination and occlusion. Although bounding boxes are the traditional way to localize subjects in videos, poses provide a richer description of the subject's bodies which can be used towards this goal (Section 1.4.3). Pose information has been used in conjunction with video to filter the information input to action detectors [124–126]. However, this has been done in other settings, and research is needed to understand the possible benefits of these methods in the mingling setting.

THE MULTI-MODALITY CHALLENGE

The multi-modality of most mingling datasets creates the challenge of leveraging those modalities for better prediction performance. Here we highlight two important challenges in the design of multimodal systems:

The fusion challenge While videos are high-dimensional, the dimensionality of modalities such as acceleration is generally orders of magnitude lower, for the same temporal window size. This poses a challenge for early fusion approaches, for these two modalities. Perhaps for this reason, feature-level and late fusion are much more common in current literature [83, 89]. In general, late and feature-level fusion approaches have been enough to observe improvements from multi-modality [80, 127, 128], but taking full advantage of certain modalities like acceleration could require *smarter*, possibly *earlier* fusion approaches.

Leveraging pre-training A lot has been attempted in machine learning literature regarding multi-modal fusion, both in deep learning architectures [129] and traditional

classifiers [130]. In deep learning applied to social signal processing in particular, it is often necessary to use pre-trained weights to achieve competitive performances, due to the relatively small size of the training datasets. This imposes an additional constraint on a multimodal architecture. Opportunities exist to research multimodal architectures that take advantage of pre-trained weights from single-modality models.

MODELLING GROUP INFORMATION

Mingling settings contain multiple simultaneous, and dynamically changing conversation groups. The social signals and experience of subjects interacting within the same group have a high degree of inter-dependence [131]. The dynamics of turn-taking in groups are highly rule-bound, with most of the time one person speaking while the interlocutors listen. This inter-dependence between social signals has been studied for many years. The phenomenon of *interpersonal synchrony*, used to refer to individuals' temporal coordination during social interaction has received considerable attention in social signal processing, with efforts to define it [132], measure it [25, 133], and relate it to higher level variables like success in cooperation [134, 135], cohesion [136], affect [66, 137], attraction [56] or relationship quality [138]. The related phenomenon of mimicry, the tendency to copy the actions of the interaction partner, has received more attention in psychological research [139], where it has been linked to similar outcomes [59, 140–142]. Most work modeling these phenomena has made use of manually engineered features designed to capture aspects of the phenomena (eg. time-lagged correlations may be used to capture mimicry).

Modeling synchrony or mimicry is only one way to utilize group information. Social signals from multiple individuals can be aggregated in many ways depending on the model's objective. Information from interlocutors, for example, can be leveraged to improve the prediction of individual behavior. Groups themselves can also be modeled and analyzed. Here, we elaborate on some of the challenges related to the use of group information in the modeling of social experience variables:

Leveraging group information Most modalities used to record social signals in mingling settings are noisy and unreliable. Video is affected by occlusion and cross-contamination. Audio is affected by background and cocktail party noise. Acceleration sensors are usually noisy. Redundancy and robustness against these factors could potentially be achieved via the use of social signals from all individuals in a conversation to predict one individual's actions or social experience, thanks to the significant degree of interdependence between interlocutors' signals and behaviors. Actions that follow rule-bound dynamics, such as speaking, or which are frequently mimicked, such as laughter, are the most likely to benefit from the use of group information. It is unclear, however, how information from multiple subjects should be modeled in a way that accounts for the underlying dynamics of group behavior. This is further complicated by the differing group sizes in a mingling setting. This has been addressed by the training of group-size-specific models. An open challenge exists in understanding whether models that learn across group sizes could be preferable.

Modelling synchrony It is hard to argue for the explicit modeling of synchrony as an intermediate step in systems whose only goal is the assessment of social experience.

In contrast to actions, synchrony is hard to interpret, hard to define precisely, and is usually considered high dimensional [132]. The study of synchrony, however, can be considered the study of social signals themselves; specifically, of the temporal inter-relatedness between signals originating from different subjects in conversation. There is, therefore, hardly a subject more central to the social signal processing field. Despite this, little is known about the most suitable way to model synchrony. This has resulted in studies making use of a variety of techniques to *capture* synchrony, and little effort to validate or compare existing methods [132]. This is not straightforward, as synchrony cannot be *annotated* and is instead assessed indirectly through outcome variables such as affect [66, 137] or attraction [56]. Furthermore, the relationship between synchrony and such outcome variables is not always clearly established. In the case of attraction, previous work shows conflicting evidence of a relationship between different types of synchrony and attraction [56]. Finally, the possible application of deep learning methods to the modeling of synchrony has received remarkably little attention [143].

Modelling group-level constructs Until now, we have talked of social experience as being inherently individual. Research, however, has also paid some attention to the modeling of *group constructs*, understood to belong to the group rather than the individual [144]. Group involvement [53] and cohesion [74] are examples. The modeling of such constructs involves bringing together information from individual subjects in a way that aligns with a definition for the construct, usually based on psychological theory.

1.5 THESIS CONTRIBUTIONS

The goal of this thesis is to contribute towards the improvement of machine assessment of social experience in interactions (Section 1.2), with an emphasis on the in-the-wild mingling sitting (Section 1.3). This involves addressing many of the limitations and challenges elaborated in Section 1.4. In the following chapters, we present the following contributions:

Exploration of the link between body acceleration and attraction (Chapter 2) . We present a study of attraction in the dyadic speed date setting. The study used accelerometer information (from chest-worn accelerometers) from 398 dyadic speed dates to analyze the relationship between body movement and self-reported affiliative goals related to attraction. The ratings, collected after a 3-minute interaction, indicated the degree to which subjects wanted to see their partner again, wanted to become friends, were romantically attracted, or were sexually attracted. We hypothesized and tested whether the mean intensity and mean changes in the intensity of a person's body movement (increase or decrease throughout the interaction) correlate to these attraction ratings. Through machine learning experiments designed to capture individual and pairwise body movement information, we investigated the predictive power of body movement information toward attraction estimation. In particular, the pairwise features used in our study were designed to capture synchrony, mimicry, and convergence information. Our work is therefore also a contribution towards understanding the predictive power of the synchrony between social signals.

1

Speaking status detection through feature filtering (Chapter 3) presents a contribution towards detecting individual speaking status in crowded scenes. Based on the observation that video shots of mingling settings contain significant cross-contamination from subjects other than the target, we present and study a model aimed at increased robustness against cross-contamination. We did this in two ways: the use of accelerometers as an alternative modality and a filtering method using body poses to exclude contaminated regions. These were incorporated into a multi-modal model via late fusion. We test the performance of uni-modal models. Through a new measure of cross-contamination obtained directly from pose information, we analyzed the model's performance at different levels of cross-contamination.

Chapters 2 and 3 were possible thanks to the considerable effort spent by researchers in collecting and annotating the datasets used for modeling and analysis. The considerations and trade-offs involved in collecting such datasets should not be underestimated (Section 1.4.2). The following two chapters of this thesis present two mingling datasets addressing challenges in the data collection stage.

ConFLab: data collection for analysis of mingling settings in the wild (Chapter 4) presents a new data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions. The dataset is unique among mingling datasets for being collected at a conference and for having been annotated for body joints. In this chapter, we introduce the design decisions and considerations around the dataset's creation, which include balancing privacy and sensor fidelity. We introduce three baseline tasks addressing key challenges in the analysis of mingling settings: body joint detection from overhead camera views, pose-based no-audio speaker detection, and F-formation detection from orientation information.

REWIND dataset: a mingling dataset with high-quality individual audio (Chapter 5) presents a mingling dataset collected during a business networking event. The particularity of this dataset with respect to previous work is that it contains high-quality individual audio recordings for 32 subjects in addition to video, pose, and acceleration data. This opens the door to the annotation of verbal behavior and the study of relationships between verbal and non-verbal behavior. (Section 1.4.2). In this chapter we present the dataset and benchmark tasks of speaking status detection from body movement (as captured by video, poses and acceleration data). We also discuss in more detail the opportunities brought about by the availability of audio.

While ConFLab is relevant for the research possibilities created by the data itself, a good part of the relevance of this project came from the methodological contributions to the data collection and annotation processes that made ConFLab possible and addressed some of the limitations in current practices (Section 1.4.2, 1.4.3). Pose annotations in a dataset the size of ConFLab were made possible by developing new, time-efficient video annotation techniques and an annotation framework for continuous annotation, the process of annotating data *while it's being watched*.

Covfee: a web software framework for continuous annotation (Chapter 6) Here we present a software framework that both a) implements novel continuous annotation techniques for keypoints and actions, used in the crowd-sourced annotation of

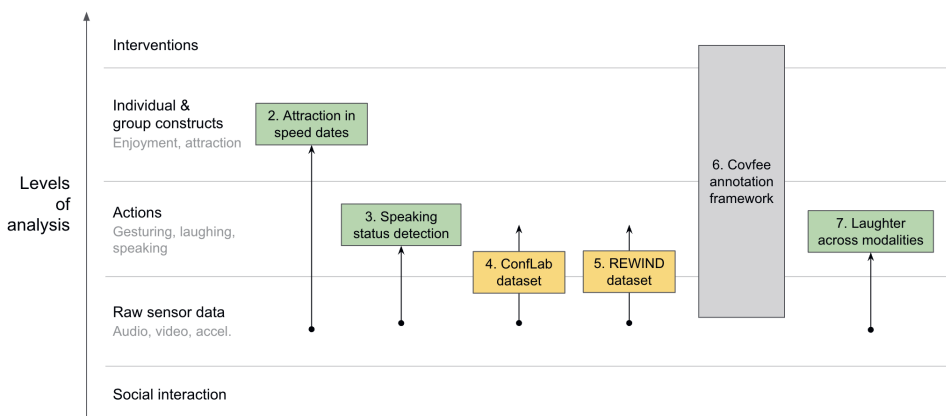
ConfLab, and b) lowers the bar for experimentation with new kinds of annotation techniques and interfaces. In addition to the framework, we present the two specific techniques used for continuous annotation of keypoints and actions in ConfLab. We present results and analysis of two case studies evaluating these techniques. In the case of continuous keypoint annotation, we ask questions about its efficacy in terms of inter-rater agreement and annotation time. The Covfee framework addresses some of the key challenges faced when annotating a human behavior dataset, and when studying annotation processes. (Section 1.4.3).

Exploration of differences in the annotation of laughter across modalities (Chapter 7)

focuses on an often-overlooked problem: the effect of the availability of different modalities during annotation. We focus on the video-only condition, due to its relevance in the annotation of in-the-wild mingling datasets. We ask whether annotations of laughter are congruent across modalities, and compare the effect that labeling modality has on machine learning model performance. We do this for models for laughter detection, intensity estimation, and segmentation, three tasks common in previous studies of laughter. Our goal is to inform the future study of laughter, through an improved understanding of the consequences of annotating laughter from a particular modality.

Figure 1.3 situates the contributions of this thesis in terms of levels of analysis of social interaction. The box indicates where the primary dependent or predicted variable of interest resides in this hierarchy. Incoming arrows indicate the source of the input or independent variables. For chapters presenting datasets (yellow) the location of the box reflect the type of data in the dataset (raw signals and annotations) and the arrow the benchmark task in the dataset paper. For Chapter 6 (Covfee annotation framework) the box reflects the levels that can be studied using the framework.

Figure 1.3: Contributions in this thesis situated with respect to the levels of analysis discussed in this chapter.



2

INDIVIDUAL AND JOINT BODY MOVEMENT ASSESSED BY WEARABLE SENSING AS A PREDICTOR OF ATTRACTION IN SPEED DATES

Interpersonal attraction is known to motivate behavioral responses in the person experiencing this subjective phenomenon. Such responses may involve the imitation of behavior, as in mirroring or mimicry of postures or gestures, which are associated with the desire to be liked by an interlocutor. Speed dating provides a unique opportunity for the study of such behavioral manifestations of interpersonal attraction through the elimination of barriers to initiating communication while maintaining significant ecological validity. In this paper, we investigate the relationship between body movement, measured via accelerometer sensors, and self-reports or ratings of attraction and affiliation in a dataset of 399 speed dates between 72 subjects. Through machine learning experiments, we found that both features derived from a single individual's body movement and features designed to measure aspects of synchrony and convergence of the couple's body movement signals were predictive of different attraction ratings. Our statistical analysis revealed that the overall increase or decrease in an individual's body movement throughout an interaction is a potential indicator of friendly intentions, possibly related to the desire to affiliate.

2.1 INTRODUCTION

Increased eye contact, smiling, laughter. It's not hard to find these behaviors portrayed as manifestations of attraction in popular culture. Research has shown that it is with good reason, as many of these behaviors, associated also with communicating trust, have been related by meta-analyses to self-reported attraction [60]. Less prevalent in popular culture but similarly researched throughout decades in social psychology are the phenomena of synchrony and mimicry as manifestations of attraction. Recently, computational social science has contributed its share of research in these areas [132].

A complete computational study of the manifestations of attraction in human behavior must necessarily encompass multiple layers, starting with the definition of the phenomenon, including the collection or procurement of suitable measurements, and the selection and interpretation of a computational model. As with many studies interested in such hypothetical constructs, subjectivity and interpersonal differences in the understanding of a phenomenon necessarily play a role in the analysis and interpretation of results. The use of machine learning models adds statistical power, normally at the expense of interpretability, and specially so for very high-dimensional data.

The advances in sensing technologies and the possibilities of sensing human behavior have brought interest in the automatic assessment of human behavior in the social signal processing community [96] originated in computer science. Many of the computational studies of attraction have been motivated by this goal. One reason is the possibility of building tools that can help people modify their behavior in their relationships via automatic feedback. Modern wearable devices make possible the measurement and provision of real-time feedback during interactions. Behavioral insights are also applicable in the development of more human-like virtual agents or robots and in science, in the development of tools that improve the time and possibly quality of psychological and sociological research.

Our line of work aims to investigate how we can automatically estimate interpersonal attraction by quantifying the body movement of the subjects involved, using wearable sensors. In a previous paper predicting the outcome of speed dates using joint body movement features [145], we have shown that it is possible to do so above chance level using features calculated using both participants' body movements. We proposed interpretable movement and coordination features inspired by previous literature that can be extracted from a single body-worn accelerometer.

In this paper, we take a broader approach by comparing, through statistical tests and machine learning experiments, the predictive power of individual body movement features (derived from a single person's movement) with that of joint movement features (derived from both people in the interaction) for the prediction of the attraction self-reports in our dataset. We hypothesize and test whether the mean intensity and mean changes in the intensity of a person's body movement (increase or decrease throughout the interaction) significantly correlate with our attraction labels. Features obtained from a single individual's movement are compared with the previously proposed joint features, designed to capture different aspects of interpersonal coordination, to assess the predictive power of individual and joint body movement. Furthermore, we significantly expand the background literature that supports our joint body movement features and test for the occurrence of convergence or divergence of body movement in our short date interactions,

to determine whether this phenomenon could be observed at all in our subjects. Finally, we performed an ablation study to understand the relative importance of the different types of features when used in isolation.

In section 2.2, we start by presenting our review of literature on attraction and interpersonal body movement phenomena including synchrony, mimicry, and convergence. In section 2.3 we present the dataset used to test our hypotheses, as well as the individual and joint body movement features proposed. Finally, we present our results and discussion on the relationship between body movement and individual attraction. We test the hypothesis that an increase or decrease in overall body movement throughout a short interaction can be related to self-reported attraction scores. In the computational stage, we used individual body movement features to directly predict the ratings of attraction. We also investigated the automatic prediction of joint attraction using *match* labels extracted from the individual ratings. In this case, we used joint features obtained from the acceleration signals of both interactants.

2.2 ATTRACTION AND BODY MOVEMENT

The following sections review works in psychology and computer science that address attraction and the phenomena of synchrony, mimicry, and convergence (with a focus on body movement) and their possible role as manifestations of attraction in face-to-face interactions.

2.2.1 INTERPERSONAL ATTRACTION

Despite the large body of work on the subject, attraction remains notoriously hard to define. The way attraction is treated in recent research does not deviate greatly from the situation in 1969 [146], where most research considers attraction as an attitude, defined as a “readiness to respond toward a particular object favorably or unfavorably”, or a “tendency or predisposition to evaluate an object in a certain way”. Attraction is thus generally conflated with a positive attitude, and the most common technique to assess an individual’s attitude remains self-report. The lack of consensus is not limited to the question of how to define and measure attraction. Montoya [147] lists several other contentious topics that have resulted in a “fragmented field, one that proceeds without a unifying theoretical model”.

Multiple works have explored the possibility of attraction as a multi-dimensional phenomenon [146–148] that cannot be summarized in a scale from negative to positive attitude. Montoya [147] present a two-dimensional model of attraction, with an affective and a behavioral component that are the consequence of an assessment of a target’s willingness and capacity to facilitate the individual’s goals and interests. The affective component reflects the “quality of one’s emotional response towards another”, while the behavioral component “reflects one’s tendency to act in a particular way toward another”. Although in many cases both components are said to align, there are occasions in which they diverge. Attraction is said to differ from love, friendship, attachment, and other related constructs in that it is an “immediate evaluation of a target person”, that characterizes interpersonal experiences in general.

Among computational studies, attraction has been conflated with interest. Gatica-Perez defines the term interest as “people’s internal states related to the degree of engagement

displayed, consciously or not, during social interaction” [149]. He also notes that this engagement may arise from different factors such as interest in the topic of a conversation, attraction to the other person, or social rapport. In this work, we make use of the terms *attraction* and *interest* interchangeably, as expressing a desire to maintain or increase contact with another person and encompassing friendly, romantic, and sexual intentions.

A big part of the work on attraction has been done in speed date settings, where self-reported attraction can be obtained from questionnaires filled in by participants [150]. Previous work investigated romantic, friendly, and business interest between partners by extracting four types of social signal measures from audio: activity, engagement, emphasis, and mirroring and successfully predicted each type of interest using these features [151]. Prosodic, dialogue, and lexical features extracted from audio recordings have also been used to predict both flirtation intention and perception [152].

Research also has explored the different mechanisms and strategies used when searching for short-term and long-term partners [153], which unsurprisingly differ between men and women. It has been noted that men tend to relax their standards further than women when seeking short-term mates and tend to have higher preferences for physical attractiveness in short-term than long-term mates [154]. Courtship behavior such as flipping of the hair and moving the shoulders has been observed more particularly in women, while men tended to cross and uncross their legs more often [155].

Previous work [55] found that positional features extracted from video such as position, distance, movement, and synchrony are indicators of attraction. Their results also indicated that separating male and female training data increased performance on the task. Cabrera-Quiros et al. attempted to classify attraction levels between participants using statistical features extracted from accelerometer data [89]. For them, separating male and female data did not improve prediction performance.

2.2.2 INDIVIDUAL BODY MOVEMENT AND ATTRACTION

Numerous factors determine our body movement during an interaction. While some of them can be related to variables accessible to measurement, like our own speech output [156, 157] or environmental stimuli like music, many are understood to be modulated by our internal states.

Although to the best of our knowledge the direct relationship between attraction and intensity of body movement has not been studied in a speed dating setting, a link between the two can be made through physiological arousal. Arousal levels have been studied as a correlate of attraction with significant results. Most studies in this area manipulate arousal via physical activity [158, 159] or by startling subjects [160], finding that increased physiological arousal resulted in higher attraction ratings when compared to baseline arousal. While these results would suggest that arousal is the cause of increased attraction, and not conversely, the direction of the relationship is not important as it relates to predictive performance.

2.2.3 SYNCHRONY, MIMICRY, CONVERGENCE AND THEIR ROLE IN ATTRACTION

The behavior of our interlocutor is another factor that influences our body movement in an interaction [139]. Numerous terms have been used in literature to refer to the dependence

in the behavioral signals of dyadic partners, such as synchrony [66, 132], mimicry [61], coordination [64, 161, 162] and chameleon effect [139].

Delaherche defines synchrony as the “dynamic and reciprocal adaptation of the temporal structure of behaviors between interactive partners”, where the important element is “the timing, rather than the nature of the behaviors” [132]. Interactional mimicry, on the other hand, has a slightly more precise definition: “when a behavior is repeated by an interaction partner within a short window of time, typically no longer than three to five seconds” [139, 163].

However, there is no consensus for the previous or any definition of synchrony. Bernieri defines coordination as “the degree to which the behaviors in an interaction are nonrandom, patterned or synchronized in both form and timing” [164], where synchrony describes the *timing* dimension. Other authors, however, have followed even less inclusive definitions. Paxton defines synchrony as a special case of coordination, where the same behavior is performed at the same time, thus conflating it with behavioral mimicry [165].

Although mimicry may be of speech, facial expressions, head movement, laughter, emotional responses, and other *observables* (ie. the behavior we observe in others) [137, 166–170], some of which cannot be easily delimited in time, we are interested in body movement mimicry, also termed *behavioral mimicry*, *behavioral matching* or *chameleon effect* [63]. This includes the repetition of the same gestures (eg. hair touching), movement of the trunk (eg. leaning forward), and the use of similar postures.

We abide by the definition by Delaherche [132], and consider mimicry to overlap with synchrony, and coordination to be an umbrella term including both phenomena and referring to all “nonrandom and patterned behaviors during a social interaction” [161, 162]. Although episodes of body movement mimicry can be considered episodes of synchrony under the definition presented, insofar as repetition of the same action implies some degree of synchrony, this repetition might be performed in a highly uncoordinated manner (eg. waiting too long or too little to reciprocate a handshake may be perceived as awkward). We consider that the measurement of the kind of coordination that facilitates social interaction requires access to contextual variables, and cannot be agnostic to the nature of the actions. Like most empirical studies, we adopt a more functional approach with measures of coordination that include aspects of both synchrony and mimicry and can be defined for behavioral time series, such as mutual information.

Synchrony has been studied especially in its link to affect, where a positive association has been found [66]. Previous work found that temporal coordination of same-sex dyads changed depending on whether they described liking, disliking, or being unacquainted with each other [171]. Synchrony has been found to relate to multiple individual outcomes like reduced anxiety and a tendency to self-identify in terms of relationships with others; as well as interpersonal outcomes like increased harmonious feelings and prosocial behavior [162]. Other studies have found that synchrony could relate to communication competence [172]; that synchrony decreases significantly during arguments [69], that more synchronous groups are perceived as more united [173] and that synchrony occurs in the psychotherapy setting [174] and could positively affect ratings of the bond with the therapist [175].

Mimicry, on the other hand, has been linked repeatedly to rapport and liking, increased mimicry leading to more favorable evaluations from an interaction partner [59] and to higher ratings of smoothness of the interaction [62]. Furthermore, having an affiliation

goal was found to increase non-conscious mimicry; and people who unsuccessfully affiliate in an interaction were found to mimic more, providing evidence for mimicry being used as a tool to achieve affiliative goals [59, 63]. Computational studies have estimated team cohesion in meeting settings using audio-visual cues and mimicry features [74, 176] with performance significantly better than random.

In the courtship setting, a meta-analysis found mimicry of nonverbal behavior to be associated with self-reported attraction [60]. In a similar context, it has been found that nonverbal mimicry is positively associated to a romantic interest in an interlocutor [61], that people who are involved in a romantic relationship mimic an attractive opposite-sex other to a lesser extent than people not in a relationship, and that they mimic less the closer they are to their current partner [39]. Beyond mimicry of nonverbals, similar associations have been found for language similarity between partners [177]. A study with speed dates [140] found that men evaluated the interaction more positively when they were mimicked by their female partner, while also increasing their ratings of the sexual attractiveness of the woman. In a study on four-minute speed dates, authors found no evidence that attraction ratings can be predicted by mimicry of certain coded behaviors (smiling, laughing, head shaking, hand gestures, face touching). The study did find evidence that synchrony in physiological signals like heart rate and skin conductance predict attraction [178]. Evidence for physiological synchrony has been found in other contexts too [179]. A more recent study [180] found that coupling in body swaying during speed dates predicted interest in a long-term relationship.

2.2.4 MEASURING SYNCHRONY, MIMICRY AND CONVERGENCE

When it comes to measuring synchrony and mimicry, it is clear that it is hard to separate these two phenomena from one another. Microanalysis from videos consists in the fine-grained coding of the timing of particular within-action moments, which can be used to measure differences in timing, related to synchrony. However, this technique is expensive in terms of human effort [64]. The coding of actions or behaviors has been prevalent in the literature as a way of quantifying action imitation or mimicry [39, 61, 140], which also enables the analysis of leading and following behaviors and roles [181]. However, behavioral coding is also expensive and cannot be used for the study of synchrony without fine-grained temporal resolution or lower-level annotations (ie. microanalysis). Therefore, many studies have resorted to the use of motion energy analysis [65, 66] from videos, wearable accelerometers or motion tracking methods [67–69]. All of these methods result in time series that act as proxies for the motion of a particular body part or as an average of body movement energy.

Multiple methods attempt to derive a synchrony measure from such time series using, for example, windowed correlations between them, possibly with different time lags [68]. It is clear, however, that correlation-based measures capture elements of both synchrony and mimicry, as both the nature and the timing of actions can affect them. The length and delay between windows are critical in this process. Schoenherr [70] compared different such time series analysis methods present in literature, including global (whole time-series) Pearson correlations and windowed correlations. The authors experimented with different ways of summarizing these outputs into scalar synchrony measures and found that these measures were only partially correlated to one another. Furthermore, they did not find

evidence of a common factor, concluding that these measures capture different aspects of synchrony.

Some recent studies using acceleration signals used cross-recurrence quantification analysis (CRQA) [71, 72]. This method allows researchers to measure the extent to which two streams of information exhibit similar patterns in time while answering questions about the characteristic time lags in the interaction [165]. Computational methods for the discovery of mimicry episodes have also been presented [73].

Datasets have been created for the study of mimicry, although in very different and specific settings like political discussions, role-playing games, and negotiations [182, 183].

The definition of the *interpersonal convergence* is somewhat more clear. We abide by its most common definition as an increase in similarity, according to some measure of similarity between features of interest [184, 185]. A study with conversations lasting between 15 and 20 minutes found evidence for the occurrence of pitch convergence and its relation to perceived attractiveness, likability, and conversation quality [186, 187]. Convergence has also been observed in the amount of laughter in a conversation [188] and the use of iconic gestures [189]. Ogata [190] coined the term *coevolution* to refer to joint changes in body movement, and found it to be more prevalent in face-to-face than in non-face-to-face interaction. A similar study used the term *synchrony* [191].

In the speech community, the related phenomenon of *entrainment*, which can be understood to include both synchrony and convergence, has been established and studied in different acoustic-prosodic features such as intensity, pitch and jitter [184, 192], as well as turn-taking features [193] and gap lengths [185] while being related to different social outcomes [194].

Synchrony relates to convergence in that it can be the object of convergence [137], that is, individuals may become more coupled in time as an interaction progresses. Convergence is not limited to synchrony, as it can affect the nature of the behaviors (i.e. mimicry) or modulate how they are performed (e.g. their intensity). In some cases such as entrainment to external stimuli [195], synchrony and convergence may be tightly linked.

Moulder [133] wrote about the importance of using surrogate data when establishing the occurrence of synchrony, to avoid observing pseudo-synchrony, the amount of spurious synchrony expected between two individuals who are not interacting. A simple surrogate data generation method may consist of calculating synchrony between non-interacting pairs to serve as a baseline or control. These ideas are necessary for studies of synchrony [65, 66] and further apply to the study of convergence.

In conclusion, there is enough evidence in previous literature to support a link between attraction and body movement, possibly mediated by the known link between mimicry and rapport. It is however unclear whether this link is limited to mimicry or if features capturing more general coordination or convergence phenomena may also be informative. The role of individual body movement in isolation as an indication of attraction to the conversational partner remains unexplored. Furthermore, previous work does not elucidate what kinds of attraction can be predicted from wearable body movement signals and little is known about gender differences in the link between overall body movement (as measured by wearables) and attraction.

2.3 DATASET AND METHODS

In our experiments, we made use of the *MatchNMingle* dataset, a multimodal and multi-sensor dataset recorded to be used in research about automatic analysis of social signals and interactions for both social and data sciences [89]. The data was collected in an indoor in-the-wild setting. It was attempted to keep the social interactions between participants as natural as possible.



Figure 2.1: Speed dating participants wearing accelerometer devices sat opposite to each other during speed dates. [89]

2.3.1 EXPERIMENT CONTEXT

The *MatchNMingle* dataset was recorded over three days in a local bar. Each day had different participants. The event started with a speed dating round where participants of the opposite sex had a three-minute date with each other, followed by a mingling event. In this study, only the data from the speed dating part of the event was used. Figure 2.1 shows several pictures of the speed daters.

Participants were recruited from a university, fitting the criteria of being single, heterosexual, and between 18 and 30 years old. A total of 92 participants attended the event, with an equal number of men and women. The majority of the participants did not know each other. Before the event, participants were asked to wear sensors around their necks to record tri-axial acceleration and proximity, as a requisite for participation. The accelerometers recorded at a frequency of 20Hz. Participants were also made aware that they were being recorded via cameras installed on a frame above the interaction area. The recorded video data is not used in this study.

After each three-minute date with a participant of the opposite sex, participants were given 1 minute to fill a booklet with a questionnaire about their date partner indicating their interest in each other. Responses for these questionnaires constitute the ground truth for the tasks in this study.

The collection of the *MatchNMingle* dataset took place over three days. 16 male and 16 female subjects participated on the first day, each involved in 14 speed dates. In the second and third days, 15 males and 15 females took part each day, with each person participating

in 15 dates. This resulted in a total of 674 speed dates. However, due to malfunctioning wearable devices, some participants did not have valid acceleration data and the data from their speed date interactions had to be discarded. From the 92 participants in the event, 72 had valid data. Furthermore, a few interactions were removed because booklet responses were unreadable. This reduced the number of speed dates in the dataset from 674 to 399. In the final dataset, each subject is present in 11.1 speed dates on average, with a minimum of 9 and a maximum of 14 speed dates for any one subject. Each of these dates became an example in our dataset.

2.3.2 DEFINING THE GROUND TRUTH

The questionnaire that participants filled out after their dates consisted of the following questions with responses on a 7-point Likert scale (low = 1, high = 7):

1. How much would you like to see this person again?
2. How would you rate this person as a potential friend?
3. How would you rate this person as a short-term sexual partner?
4. How would you rate this person as a long-term romantic partner?

These questions were chosen because, in line with a general notion of attraction as *interest in the interlocutor* in a goal-oriented manner, they cover the most common ways in which subjects may be interested in each other in the context of an informal speed date. Concretely, the first question captures a general notion of interest by wanting to see the other person at least one more time. This interest could be towards any of the three goals implicit in the next three questions. Question 2 explicitly asks for interest in a friendship. This type of interest has been linked to rapport, with it incorporating feelings of friendship and caring, and the notion of being in sync [196]. Romantic and Sexual ratings, on the other hand, are directly related to partner choice, where a range of factors like similarity, reciprocity, physical attractiveness, and security offered by the partner are known to play a role in the assessments [197].

In Figure 2.2, we show the correlations between the raw Likert-scale ratings of the same interaction, where the goal was to understand overall gender-related differences in the way males and females treated the ratings, given that large gender-based differences in partner choice are reported in literature [197]. The first plot shows correlation between the four different ratings (questions) given by males for the same interaction; the second between ratings given by females, and the third between the ratings of the males and the ratings of the females (ie. a positive value means that men and women tended to agree in their ratings of how much they liked each other; a negative value that ratings were often opposite). Males made a big distinction between the Friendly label and the rest of the labels, but SeeAgain, Romantic, and Sexual have similarly higher levels of correlation. On the other hand, females tended to form two clusters, with Friendly and SeeAgain ratings being one (labeled similarly) and Romantic and Sexual labels being another.

Correlations between male and female responses are low, highlighting the importance of analyzing attraction first as an individual construct, as there is seldom agreement on attraction. Interestingly, only correlations involving the Sexual rating were significant.

Male *Sexual* ratings correlate negatively with all female ratings except for the Friendly intention. For females, their *Sexual* ratings correlate negatively with male *Sexual* and *Friendly* ratings.

2

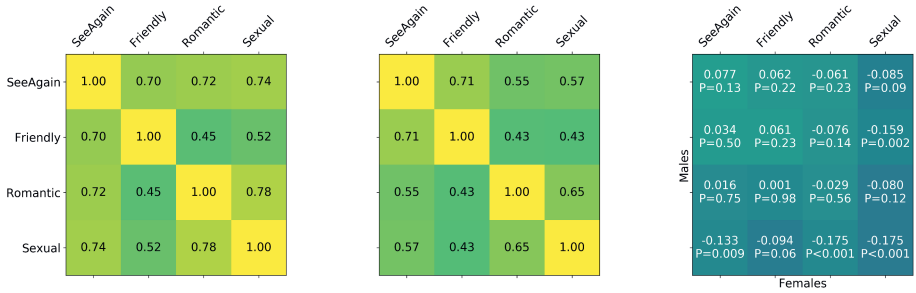


Figure 2.2: Spearman correlations of speed date responses (Likert scale from 1 to 7). (Left): Male ratings. (Center): Female ratings. (Right): Male and female ratings.

Each of these ratings was used to define different tasks for the interest prediction problem as *See Again*, *Friendly*, *Sexual* or *Romantic*, which consist in predicting the corresponding label. For a more straightforward interpretation of the results, we treated the problem as binary classification. Responses to one question were binarized by assigning a positive label to the ratings equal to or above the median (per gender) of all ratings given for that question, and a negative label otherwise. In other words, the median of the ratings was used as the threshold for binarization. The threshold was different per gender because in the experiments we also predicted separately for males and females and the distributions of scores were very different between them. Figure 2.3 shows the distribution of booklet responses and the corresponding median thresholds used for binarization. Additionally, interactions were labeled as a *match* when both speed daters had a positive label for the interaction.

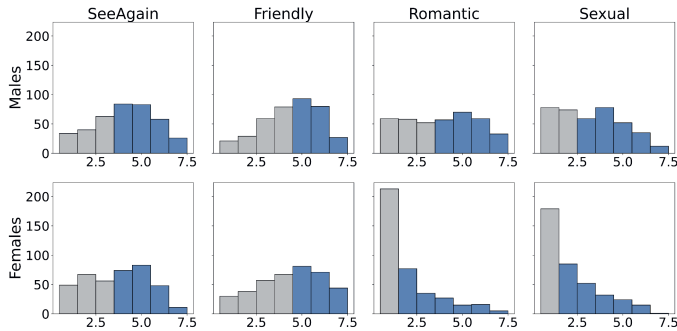


Figure 2.3: Distribution of the speed date responses for the four questions asked, for the 399 interactions in the dataset.

2.3.3 FEATURE EXTRACTION

Our method aims to model the coordination of behavior between two people in an interaction using nonverbal behavioral features extracted from accelerometer readings. We describe the feature extraction process in detail below.

PREPROCESSING

The accelerometer data consists of 3-dimensional readings recorded at 20 Hz with the X-axis capturing the left-right movements; the Y-axis up-down movements and Z axis forward-backward movements. Initially, each axis of each person's recordings is normalized by subtracting its mean and dividing by its standard deviation. This is done to reduce the effect of gravity and interpersonal differences of movement intensity in the sensor readings, and follows previous work [198]. These three normalized signals are augmented with the absolute value signal of each axis, and the magnitude of the acceleration, computed as $\sqrt{(x^2 + y^2 + z^2)}$ for a total of 7 signals.

Each of these 7 signals was divided into n -second windows using a sliding-window approach, with $n/2$ second shifts between each window. Since the optimal window size to capture relevant behavior is unknown, we chose to extract windows for multiple values of n : 1, 3, 5, and 10 seconds; all of which are included.

Similar to [198], statistical (mean, variance) and spectral (power spectral density) features are extracted from each window. Power spectral density (PSD) per window is computed using 6 logarithmically-spaced bins between 0-10 Hz, to increase the resolution at low frequencies, which contain most of the energy of human movement.

Each bin gives information about the characteristics of the person's behavior at that time window. Therefore, each bin is treated as a single feature. Combining these features results in 8 feature dimensions per window.

Computing these 8 features for each of the 7 signal types and for 4 different window sizes results in 224 low-level signals that are further used to extract behavioral coordination features, detailed in the following subsection.

An illustration of the pre-processing steps is shown in Figure 2.4.

The aforementioned low-level signals are used to extract more complex body movement features that are grouped into two categories: individual and pairwise features.

INDIVIDUAL FEATURES

For experiments using the body movement of a single individual as input, we used two simple features that quantify how low-level body movement signals change during the interaction.

Time-correlation One time-correlation feature was computed as the Pearson correlation coefficient (Pearson's r) computed between one of the low-level signals (eg. PSD bin 3 of the X axis) and time. These capture the general direction of change of the low-level signal throughout the interaction. For example, a positive coefficient for the mean magnitude of acceleration would indicate an increase in body movement intensity throughout the interaction.

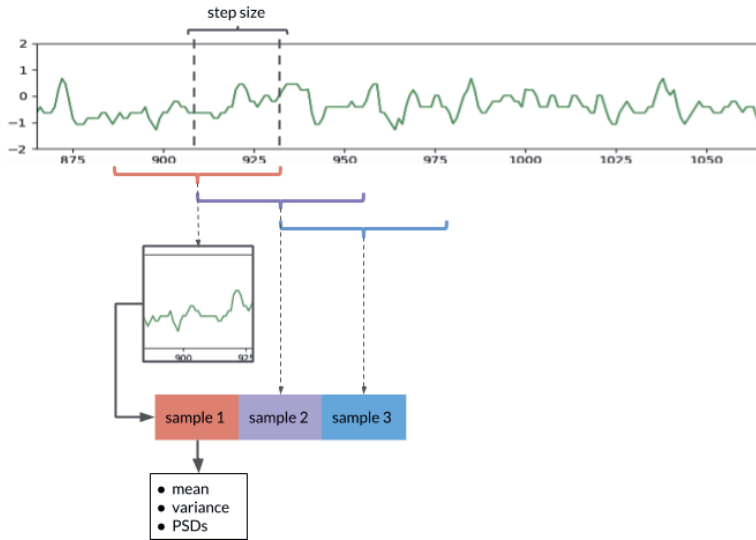


Figure 2.4: Pre-processing step: Using a sliding window approach, the signal is divided into samples from which the statistical and spectral features are extracted.

Split difference One split difference feature was computed as the difference between the mean of the low-level signal in the last third and the first third of the interaction. These features similarly capture changes in the underlying low-level signals, by comparing them at the beginning and end of the interaction.

PAIRWISE FEATURES

The following measures aim to quantify body movement behavior between two subjects. The first three measures were created to capture different types of coordination between the movement of the two people in the dyad, especially synchrony and convergence. The next two features were designed to measure convergence (or divergence), the tendency of body movement to become more or less similar during the interaction. Note that, as for the individual features, all of the following joint features are computed for the 224 multi-scale low-level signals (see Section 2.3.3. When present in this section, X and Y refer to a corresponding low-level signal (eg. the mean of the X axis of acceleration, calculated using a sliding window of 3 seconds); X for one subject, and Y for the other subject in the interaction.

Correlation Linear correlation scores have been used in the literature as a measure of similarity of overall body motion as well as motion of specific body parts such as the hands or head of two people [66, 138, 199–201].

The linear correlation between two people's body movement signals is expected to result in a score closer to 1.0 the more similar the movement of the two people, hence capturing mimicry in particular but also being affected by the precise timing of the behavior.

Correlation with a time lag has also been used to measure the linear relationship between a follower and a leader's movement [199, 200]. The following computes the correlation between X and Y signals at a positive lag of τ samples:

$$\rho_{xy} = \frac{\sum_{i=1}^{N-\tau} (x_i - \mu_x)(y_{i+\tau} - \mu_y)}{\sigma(X)\sigma(Y)} \quad (2.1)$$

where x_i and y_i are corresponding samples, μ_x and μ_y the means of the signals and $\sigma(X)$ is the standard deviation of X .

Using time lags enables capturing the leader-follower relationship of two people in a conversation. In an example case of measuring the correlation between persons A and B's movement, if a higher score is obtained when person B's signal is positively lagged, this indicates that person B is leading the interaction.

Following the literature, we use +/- 1 time step lags, and no lag for direct correlations.

Distance This movement similarity measure is inspired by the work of Nanninga [74] and adapted for movement data.

The goal is to capture when one person imitates their partner's behavior. Figure 2.5 illustrates how this feature is computed. Each sample window of Person A's signal is compared with the consecutive window of Person B's signal. To compare these windows, the distance between low-level features of these windows is computed, resulting in distance scores $D = [d_0, d_1, \dots, d_n]$ for the entire interaction.

From these distance scores, minimum ($\min(D)$), maximum ($\max(D)$), mean ($\text{mean}(D)$) and variance ($\text{var}(D)$) are computed and used as features. Since this feature is asymmetrical, the reverse is also computed.

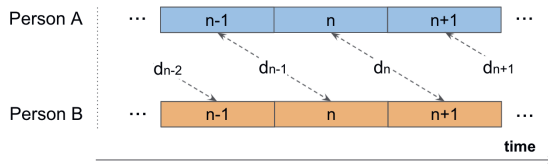


Figure 2.5: Distance features. Each time sample is compared with the other signal's preceding time sample.

(Normalized) Mutual Information Mutual information computed between the random variables corresponding to two movement signals has also been used in the literature to capture the dependence between two people's behavior [198, 202]. In our case, it captures the dependence of two people's behavior on each other. It quantifies how much information can be obtained about one variable by observing the other variable. Mutual information is calculated as follows:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (2.2)$$

where $H(X)$ and $H(Y)$ represent the entropy of random variables X and Y and $H(X, Y)$ represents their joint entropy. As the calculation of entropy requires knowledge of the

underlying probability distributions, we approximated $P(X)$, $P(Y)$ and $P(X, Y)$ using categorical distributions by calculating 10 bin histograms for the marginal distributions, and a 10×10 histogram for the joint distribution.

Additionally, normalized mutual information is computed by dividing by $\sqrt{H(X)H(Y)}$ to obtain a score between 0 and 1. A higher score is expected when two people have an influence on each other's behavior.

While the three previous features attempt to measure elements of coordination, the next two sections describe features that aim to capture the degree of convergence or divergence of body characteristics during the short interaction.

Time-correlation Time-correlation features try to capture if the difference between two people's behavior increases or decreases over time [74, 186]. In order to compute it, the corresponding windows of two participants' signals are compared with each other. To measure the similarity at each time step, the distance between these corresponding samples' low-level features is computed as illustrated in Figure 2.6, resulting in distance scores $D = [d_1, d_2, \dots, d_n]$, for each sample. After that, the correlation of these scores with time is computed to understand if they increase or decrease using Pearson correlation formula (eq. 2.1), and a correlation coefficient is obtained. Since the goal is to capture convergence, a decreasing distance indicates converging behavior. Therefore, the correlation coefficient is expected to be more negative for converging interactions where participants show similar behavior over the interaction.

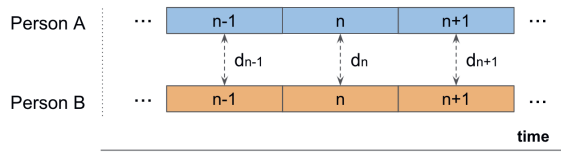


Figure 2.6: Time-correlation feature. Each time sample is compared with the other person's corresponding time sample. These distance scores are further correlated with time to extract a convergence score.

We further incorporated a second type of time-correlation feature inspired in previous work [74], where they were found to be effective at measuring para-linguistic mimicry in meetings. In this case, the first two minutes of the date interaction are taken as a *learning period* in which the baseline level of one participant is modeled and the last minute of the second participant (analysis window) is compared to this learned baseline. To understand if the second person's behavior converges to the behavior exhibited by the first person during the learning period, the N low-level features in the analysis window are compared to the learning period's low-level features. We compared features by subtracting their means, resulting in distance scores $D = [d_1, d_2, \dots, d_N]$, for each window in the last one minute of interaction as illustrated in Figure 2.7. The correlation of scores D with time was then computed using Pearson correlation. A negative correlation coefficient indicates a behavior that becomes more similar to the other person's baseline. Since this feature is asymmetrical, it was computed for both possible combinations.

The rationale for including these features is the capturing of a baseline level of body movement of one participant for a long period (the 2 min *learning period*) compared to

other features (which compare individual windows) to measure the tendency of the other participant to approach or reject this baseline level.

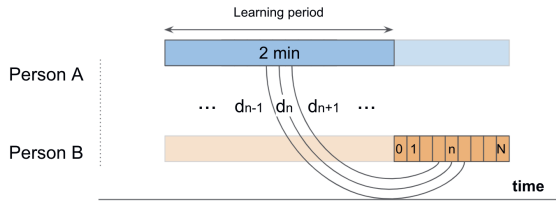


Figure 2.7: Convergence features with a learning period. Each window in the last 1-minute period was compared with the other person's first 2 minutes by computing a distance score between mean sample features.

Split-difference Split-difference features are inspired by the work of [186]. The idea is to measure the similarity of two people's behavior at the start and end of their date interaction and compare these similarities. It is expected that behavior will be more similar at the end of the interaction if convergence occurs. To capture this, the first and second half of the signals are taken as illustrated in figure 2.8. The similarity d_0 between the first half's features of the two persons is computed. An equivalent similarity d_1 is calculated for the second half. One feature corresponds to the difference between these similarities: $c = d_1 - d_0$. This difference is expected to be negative when convergence occurs.

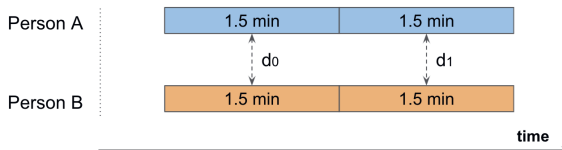


Figure 2.8: Split-difference feature. The difference between both persons' features is computed for each half of the interaction.

Table 2.1 summarizes all the features used in our experiments, along with their dimensionality. Joint features are separated into those measuring coordination and those measuring convergence of behavior as explained in this section.

2.3.4 DIMENSIONALITY REDUCTION

After extracting the features, they were processed to reduce the dimensionality of the feature space. We applied principal component analysis (PCA) and the top principal components preserving 95% of the variance were kept. Features were then normalized to have zero mean and unit standard deviation, as is standard practice for classification.

Table 2.1: Summary of the individual and joint features used to predict attraction ratings. *Total* indicates the size of each feature vector or number of individual features.

Type	Category	Feature type	Total
Indiv.	-	Time Correlation [Sec. 2.3.3]	224
		Split-difference [Sec. 2.3.3]	224
Joint	Coordination	Correlation [Sec. 2.3.3]	672
		Distance [Sec. 2.3.3]	1792
		Mutual Information [Sec. 2.3.3]	336
	Convergence	Time Correlation [Sec. 2.3.3]	784
		Split-difference [Sec. 2.3.3]	224

2.4 RESULTS

Our experiments can be separated into three parts. First, we investigate the relationship between body movement intensity and attraction at the individual level through a correlation analysis. Second, we attempt the automatic prediction of the individual binary attraction levels using a set of convergence features extracted only from individual body movements. Finally, we investigate the automatic prediction of the mutual attraction labels using features designed to capture synchrony and convergence, thus derived from both individuals' time series during these interactions.

2.4.1 BODY MOVEMENT AND ATTRACTION

We start by investigating the relationship between overall body motion and attraction, starting with a simple hypothesis: the intensity of overall body motion in the interaction is linked to attraction. The magnitude of the accelerometer signal (see 2.3.3) was normalized per participant by dividing by the participant's mean magnitude over all its interactions. This is expected to capture relative changes in individual body movement and remove interpersonal differences in body motion energy.

Table 2.2 shows the results of correlating the average intensity of the accelerometer readings with the questionnaire responses (7-point scale) for males and females separately. Spearman's r was used to avoid excessive influence from individuals with extreme body movement energies. No significant correlations were found, and in fact all correlation coefficients were negative, suggesting a weak opposite relation.

For the previous calculations, body movement energy was averaged for an interaction, meaning that we did not capture the effect of the interaction on the body movement intensity of participants (increasing or decreasing). Our next hypothesis tests whether net increases in body movement indicate heightened interest, possibly through an increasingly animated conversation. To quantify this we calculated correlations between body movement intensity throughout the interaction and time. Most correlations were significant ($\alpha = 0.05$), indicating a substantial change in body movement throughout the interaction. For females, from the total of 398 interactions, 32 interactions had a significant increase, and 204 had a significant decrease in body movement. For males, 43 coefficients were positive, and 198 were negative. The fact that participants were seated and changed seats

Table 2.2: Correlations between mean intensity of body movement and the attraction ratings did not produce any evidence of increased or decreased body movement being linked to attraction.

Attraction	Males		Females	
	Spearman's r	p-value	Spearman's r	p-value
SeeAgain	-0.041	.41	-0.098	.05
Friendly	-0.005	.92	-0.050	.32
Romantic	-0.062	.21	-0.022	.66
Sexual	-0.077	.12	-0.057	.26

between interactions is the most likely cause of the high number of interactions with decreasing movement intensity. Even though the analyzed interactions start a few seconds after participants have seated and greeted each other, it is possible that this moment of higher arousal influences the rest of the interaction, and that participants take more time to reach a state closer to their baseline. The same is not true for the end of the interaction. The recording ended right before a bell rang during the event, indicating participants to switch partners.

We used these correlation coefficients as a variable quantifying the effect of the interaction in body movement. Table 2.3 shows the results of correlations between corresponding r values and speed date responses. In this case three of the correlations were found significant. Interestingly, for all labels correlations are positive for males and negative for females. The strongest significance was found for the *Friendly* and *SeeAgain* labels for both males and females. A possible explanation for this last fact is that high rapport drives these changes in overall body movement. A stronger link of high rapport to the *Friendly* ratings, in comparison with *Sexual* and *Romantic* ratings where other aspects like physical attractiveness play a big role, would explain the differences in significance. *SeeAgain* ratings are inherently more ambiguous and the analysis of section 2.3.2 indicates that males and females tended towards different interpretations. Note however that all coefficients are below 0.5. The rapport link would imply that high rapport is associated with increases in male body movement (or less steep decreases given that most of the r values were negative) and with stronger decreases in female body movement throughout the interaction.

Table 2.3: Correlations between the individual time-correlation scores and attraction labels. An asterisk (*) marks significant correlations ($\alpha = 0.05$)

Attraction	Males		Females	
	Spearman's r	p-value	Spearman's r	p-value
SeeAgain	0.084	.093	-0.107	*.032
Friendly	0.106	*.035	-0.112	*.026
Romantic	0.068	.18	-0.047	.35
Sexual	0.078	.12	-0.034	.49

AUTOMATIC PREDICTION OF INDIVIDUAL INTEREST

We predicted individual interest based on an individual’s accelerometer features (extracted as per section 2.3.3) and the joint movement features extracted from both speed daters. In these experiments, we train a classifier to predict attraction from male to female and from female to male. A logistic regressor (linear model) with L2 regularization was chosen as the classifier for the task.

The model was evaluated via 10-fold cross-validation. To avoid having dates from the same subject in train and test sets, the cross-validation split was done via a leave-n-subjects-out approach. When male labels are predicted, the dates from a number of males (three subjects for most folds) are separated as test set in such a way that their dates are not present in the training set. The equivalent happens when female labels are predicted. A nested cross-validation loop within each fold was used to tune the regularization parameter. To obtain a measure that is unaffected by the class imbalance, the Area under the Receiver Operator Characteristic (ROC-AUC) was used as the performance measure.

Performances for different attraction type predictions were compared to a random baseline classifier (expected AUC of 0.5), via a statistical test on the 100 classification scores obtained from running 10-fold cross-validation 10 times (10x10-fold cross-validation). P-values were obtained by using the correction to the paired Student t-test initially proposed by Nadeau and Bengio [203] and recommended [204] for enhancing replicability of the p-values obtained from 10x10-fold cross validation classifier scores. Obtained results are shown in Table 2.4. Note that AUC scores lower or equal to 0.5 indicate that the classifier could not discriminate between the two classes above chance.

Table 2.4: Mean AUC scores obtained in individual interest prediction tasks via 10x10-fold cross-validation. P-values are for the probability of observing more extreme cross-validation scores under a true mean of 0.5 AUC, calculated using the Nadeau and Bengio correction to the paired Student-t test for comparing classifiers [203].

	Individual Features				Joint Features			
	Males		Females		Males		Females	
Attraction	AUC	p-value	AUC	p-value	AUC	p-value	AUC	p-value
SeeAgain	0.482	.35	0.588	*.008	0.508	.73	0.584	*.012
Friendly	0.482	.27	0.555	.06	0.510	.76	0.608	*.0002
Romantic	0.493	.71	0.483	.22	0.601	*.005	0.519	.49
Sexual	0.501	.97	0.574	*.011	0.573	.06	0.531	.34

2.4.2 JOINT BODY MOVEMENT AND ATTRACTION

This section focuses on joint movement measures (calculated from both subjects’ movement signals) and their relation with mutual ratings of attraction. As before, this is done through both statistical results and classification experiments.

CONVERGENCE OF BODY MOVEMENT

Following previous literature which explored the phenomenon of convergence in features of speech in dyadic conversation [186] we investigated whether we can find evidence

of convergence of body movement between interacting partners. Previous work found important evidence that several pitch features converge globally throughout a conversation, independent of the perceived attractiveness or likability of the interlocutor.

We hypothesized that during the 3-minute dates, the participants' movement characteristics converge or diverge due to the effect of the social interaction. To test our hypothesis we compared the convergence scores of interacting and non-interacting pairs. We created non-interacting feature pairs by randomly matching input signals from males to females who were not conversing together. Convergence scores were calculated for real and artificial non-interacting pairs as described in section 2.3.3. However, for these experiments we used only the time-correlation and split-difference convergence features due to their easy interpretation and because they capture the complete temporal extent of the interaction. We used only the convergence features extracted using windows of 3 seconds because since convergence features are correlations with time, scores for different window sizes are expected to be highly correlated.

It was clear however that there is no significant difference in convergence of body movement magnitude. Not only did we find no significant difference between the means of interacting and non-interacting pairs ($P = 0.97$), but more significantly converging and diverging interactions were found for randomly matched pairs than in the actual interaction. Most of the convergence behavior can thus be attributed to an overall reduction in body movement rather than to the effect of the social interaction.

Given these results, we performed a more complete analysis by using similar one-tailed t-tests ($\alpha = 0.05$) for the rest of the time-correlation and split-difference convergence features, this time for all the Power Spectral Density bins, and variance. However, from the total of 112 features only three of these tests were significant, less than expected by chance. We found thus no evidence of a difference in the mean of convergence features between interacting and non-interacting pairs.

AUTOMATIC PREDICTION OF MUTUAL ATTRACTION

In these experiments, we train a classifier to predict the mutual attraction or *match* binary labels using the joint movement features presented in section 2.3.3. The goal is to test the predictive power of body movement in interactions where both participants rated each other above average in a particular item. Note that because *match* labels were obtained as the intersection (logical *and*) of the individual labels (Section 2.3.2), the dataset is more unbalanced in these tasks. Furthermore, because *match* labels come from both subjects, we did not perform cross-validation splits at the person level. Instead, we used a traditional split at the example (speed date) level.

Table 2.5: Mean AUC scores from 10x10-fold cross-validation for mutual interest prediction tasks. P-values are for the right tail of the t-distribution. A random classifier has an AUC of 0.5.

Label	AUC	p-value
See-Again	0.553 (0.011)	.06
Friendly	0.562 (0.011)	*.02
Romantic	0.495 (0.016)	.88
Sexual	0.551 (0.015)	.12

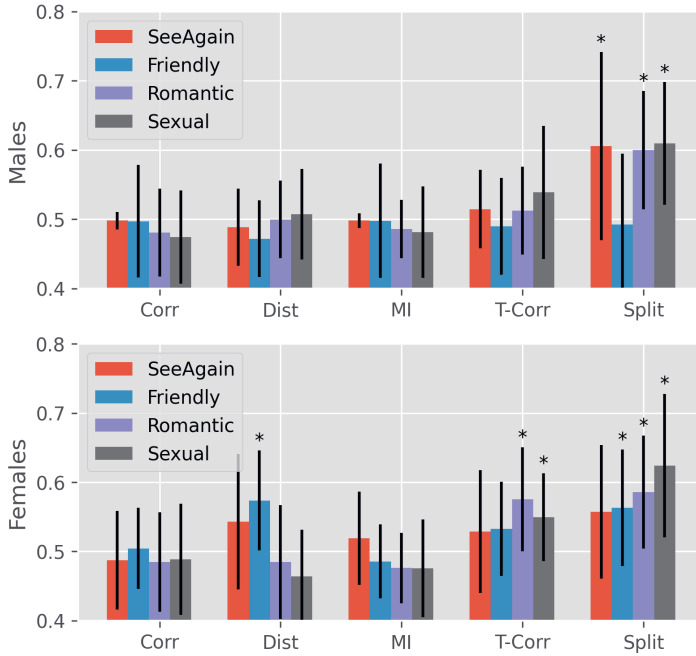


Figure 2.9: Results of the ablation study for individual interest prediction tasks with different sets of features. The bars indicate the mean and standard deviation of the AUC scores from 10x10-fold cross-validation. An asterisk indicates performance significantly better than the random baseline classifier.

As before, we use a logistic regressor with L2 regularization trained and evaluated via 10x10-fold cross-validation using the AUC score as the evaluation metric. Obtained classification scores are shown in Table 2.5. In this case, although three of the mean scores are above 0.55, more than in the individual tasks for males and females, only predictions of the *Friendly* labels reached significance and *Romantic* attraction was the hardest to predict. We found thus no clear evidence that predicting matches in this way can be done with better performance. This could suggest that mutual attraction is less characteristically expressed in body movement. However, the lower performance could partly come from the lower number of positive labels (25% on average). Imbalance is however hard to avoid as it is a feature of the interactions themselves, where matches are much more rare than one-sided attraction. Experiments with balanced class weights in our Logistic Regressor, a technique that can offset class imbalance, delivered performance statistically indistinguishable from the results in Table 2.5 for all four labels.

2.4.3 ABLATION STUDY: FEATURE TYPE IMPORTANCE

In this section, we present the results of an ablation study to understand the relative importance of the different types of features (Table 2.1) in our method. The goal is to understand how different sets of engineered features affected the results in previous sections. We focus on individual interest prediction using joint features, where we had 5 different

feature sets designed to capture different aspects of coordination.

The results of the ablation study are shown in Figure 2.9. The experimental setup and evaluation were the same as detailed in section 2.4.1. It stands out from these results that convergence-related features (time-correlation and split difference) were in general the most relevant. These results indicate that features capturing synchrony and mimicry were barely predictive of attraction in isolation, and for males in particular, these coordination features held no discriminative power. Note that it is still possible that interactions between features are discriminative, but we limited the ablation study to the individual feature sets.

2.5 DISCUSSION

Our experiments with individual body movement revealed (Table 2.4) that the attraction of a participant can be predicted only by their movement features, with performance significantly better than random guessing. These results suggest that female attraction is more easily revealed by their body movement than male attraction. The statistical analysis (Table 2.3) suggests a possible explanation: although we found no significant correlation between average acceleration intensity and attraction, women were found to significantly decrease their body movement the more positively they rated their interaction partner in the *SeeAgain* and *Friendly* categories. For men, all correlation coefficients were positive, which suggests that an increase or a less steep decrease in body movement reveals heightened attraction. This relation, opposite to that of females, was only significant for the *Friendly* rating.

Experiments with joint features designed to capture aspects of synchrony and convergence resulted in better performance in the prediction of individual attraction. Our results indicate that performance in the detection of attraction depends not only on the type of attraction but also on the gender of the target subject. In general, for females, we found stronger evidence that *SeeAgain* and *Friendly* ratings were linked to body movement, and less so for *Sexual* and *Romantic* ratings. For males, the opposite was true in the case of joint body movement (*Romantic* and *Sexual* labels were the better predicted). This separation cuts along the distinction made by participants in their ratings (Figure 2.2). Males made a big distinction between the *Friendly* ratings and the rest of the ratings, but *SeeAgain*, *Romantic*, and *Sexual* have similarly higher levels of correlation. Females, on the other hand, tended to form two clusters, with *Friendly* and *SeeAgain* ratings being one (labeled similarly) and *Romantic* and *Sexual* ratings being another.

Different interaction dynamics likely play a role in explaining these general trends. Our results suggest that interactions in which the female seeks friendship or the male seeks romantic or sexual goals have a characteristic signature in body movement. This could be caused by the interested participant (or both of them) making an effort to affiliate with their partner. Body movement phenomena like mimicry are known to be effective tools for seeking affiliation and increasing rapport [59, 139].

The better performance of joint features compared to individual ones in predicting individual attraction indicates again that individual experience of attraction has a strong manifestation in the joint interaction, although this general trend could be a result of our particular choice of features.

In attempting to understand the relative importance of the many joint features that we used, the ablation study of Section 2.4.3 showed convergence features to be the most

important, indicating that mimicry and synchrony are less relevant to attraction compared to the less dynamic convergence features. This may appear odd in the light of the results of section 2.4.2 which established that there was no evidence of convergence taking place above chance levels. However, changes in overall body movement levels, or interactions between them captured by the classifier may hold the discriminative power. The statistical results of section 2.4.2 only show that the dyads in our dataset did not converge more often than expected by chance.

Prediction of mutual attraction delivered results significantly better than random for the *Friendly* labels (Table 2.5). Note that mutual labels have a logical relation to individual labels in that they must both be positive for a positive mutual or *match* label. Therefore, the fact that *Friendly* scores in the joint tasks are between the (low) scores obtained individually for males and the high scores obtained for females (Table 2.4), would seem to suggest that cases of one-sided female friendliness are easier to detect than when such friendly intentions are mutual. We think however that there is not enough evidence to reach this conclusion, since the greater data imbalance in the mutual tasks could explain having lower results in the mutual tasks.

The fact that no significant difference in convergence could be observed between interacting and non-interacting pairs could be an indication that convergence in overall body movement does not occur over these short timespans, or is much weaker than other factors like the significant average decrease in body movement that we measured during most interactions. However, this evidence is far from conclusive given the simplicity of the sensing modality, which only has access to the acceleration of a single body part (the chest), and is limited to a setting where participants are seated. Another possibility is that convergence manifests itself as an increase in the time-synchrony of behavior (ie. is tightly linked to synchrony), and not in the intensity or style of the movements. This would not be captured by the Time-correlation and Split-difference features, which perform a rough aggregation over the complete interaction.

An analysis directly correlating different joint features with the label of each task revealed that the types of features with the highest correlation coefficients vary with different tasks. Correlation features computed over the Z-axis were found to be often negatively correlated with *Friendly* attraction as opposed to the expectation of positive correlation that would indicate mimicry. Because the Z-axis of the accelerometers captured the forward-backward acceleration of the body, low feature values can be produced by a person's backward and partner's forward movement occurring simultaneously. This could indicate that a different kind of synchrony is at play. On the other hand, most of the correlation features extracted from PSD bins had significant positive correlations with the *Friendly* and *Sexual* attraction ratings, indicating that coupling in the frequency of movement could be a correlate of these ratings.

It was also found that *Mutual Information* features tended to have a high positive correlation with only the *SeeAgain* and *Friendly* labels whereas the *Mimicry* features correlated more often with the *Romantic* and *Sexual* tasks, offering a possible explanation for the differences in the computational results.

The fact that we found no common features correlating significantly across all of the four ratings tends to indicate that different types of attraction manifest in different behavioral characteristics.

In conclusion, our computational analysis showed that it is possible to predict speed date ratings and the derived matches using individual and joint behavioral coordination features derived from a single body-worn accelerometer. Features engineered to capture synchrony and convergence characteristics succeeded in predicting three of the mutual attraction levels and distinct individual attraction labels for males and females. Our results indicate that subtle social manifestations of attraction can be captured by wearable devices. This calls for similar studies using more complete body movement sensing. More complex wearable sensors, however, risk interfering with the interactions or limiting body movement. Alternative setups such as video recordings followed by joint position estimation algorithms are worth consideration.

Another limitation of our study is the treatment of the labels since the combination of the ratings of both partners can have a large effect on the dynamics of the interaction. An interaction where both partners have friendly intentions, for example, can be very different from one where one of them has sexual intentions instead. Not looking at the interaction between labels can therefore be limiting. However, classifying label combinations rather than single labels is impractical with our relatively small dataset.

The development of the computational study of phenomena such as synchrony and convergence via proxies, and their relation with constructs like attraction faces the fundamental problem of lack of suitable, large-scale, ecologically valid datasets. The dataset used in this study is a step in the right direction but larger wearable sensing or video datasets would allow us to more conclusively answer questions related to interpersonal gender, age, and culture-related differences in the manifestation of attraction.

ACKNOWLEDGEMENTS

This research is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

3

NO-AUDIO SPEAKING STATUS DETECTION IN CROWDED SETTINGS VIA VISUAL POSE-BASED FILTERING AND WEARABLE ACCELERATION

Recognizing who is speaking in a crowded scene is a key challenge towards understanding the social interactions within it. Detecting speaking status from body movement alone opens the door to analyzing social scenes in which personal audio is not obtainable. Video and wearable sensors make it possible to recognize speaking in an unobtrusive, privacy-preserving way. When considering the video modality, in such action recognition problems, a bounding box is traditionally used to localize and segment out the target subject, to then recognize the action taking place within it. However, cross-contamination, occlusion, and the articulated nature of the human body make this approach challenging in a crowded scene. Here, we leverage articulated body poses for subject localization, and in the subsequent speech detection stage. We show that selecting local features around pose keypoints increases generalization performance while significantly reducing the number of local features considered, making for a more efficient method. We investigate the role of cross-contamination in this effect. We additionally use acceleration measured through wearable sensors for the same task and present a multimodal approach combining both methods.



Figure 3.1: Example of a free-standing crowded scene from our dataset.

3.1 INTRODUCTION

Detection of speaking activity in free-standing social settings is a core need in building systems capable of detecting and understanding the social interactions developing in a scene. The analysis of a complex conversational scene where dozens of people stand, walk, form groups, and converse freely (see figure 3.1) is of interest in the fields of computational social science and social signal processing. Speaking status is key because of its utility in downstream tasks, where speaking predictions can be used, for example, in the quantification of individual and group measures of conversation quality like involvement [52], satisfaction [77] or affect [76], and in the forecasting of future events like speaking, gesturing and changes in position and orientation [131].

Analyzing unconstrained social scenes requires systems capable of efficient speaking status detection, due to the large number of people in the scene and the extent of natural human interactions. Although recording audio from the participants is the obvious choice for measuring speaking status, this modality is especially hard to acquire in a crowded conversational scene. Obtaining consent from participants can be challenging due to privacy concerns (from access to the content of conversations), personal discomfort from wearing a microphone, or concerns about the social acceptability of the device [205]. Furthermore, high-quality audio recording equipment can be hard to scale to study large crowds. Low-cost wearable devices like sociometric badges [88] have been proposed as a more cost-effective alternative. However, these devices suffer from audio quality issues stemming from the noisy setting, the low quality, and the omnidirectionality of the microphones [88], while still requiring subjects to wear an audio-recording device. These limitations have caused that most datasets created for the study of free-standing crowds do not contain personal audio, nor information about the speaking status of subjects [89].

The possibility of detecting speaking from body movement alone, without access to the

audio, offers a privacy-sensitive solution to these problems. It has long been established that hand and head gestures frequently synchronize with speech [156, 206, 207] while being salient cues with recognizable motion characteristics across people. While the characteristics of speech-related gestures are to some extent modulated by culture, the link between speech and gesture appears universally early in life and there are no reports of a culture where a tight link does not exist [42]. Video cameras are a convenient way to observe these gestures but their visibility is affected by factors such as occlusion and cross-contamination. We consider cross-contamination to include any case where the bodies of other people in the scene are visible within the bounding box, or area considered by the recognition system and occlusion when parts of the body of the target subject are occluded, possibly by their own body [89]. These factors are made worse when the subject of interest is localized using a bounding box [89] due to the difficulty of accurately segmenting a person’s body. Cross-contamination is especially problematic when it comes from the target’s interlocutor(s), as they are likely to be speaking when the target is listening, pushing the prediction towards a false positive.

Given recent significant advances in visual pose estimation [97, 208, 209] it is natural to use pose information to alleviate these issues. Accurate poses allow for more precise localization and filtering of the information that is input to the recognition stage. This work concentrates on using pose in a speaking status detection system. We focus on the classification of short windows of interaction using a state-of-the-art feature aggregation pipeline and explore how pose can be used to efficiently filter local features, resulting in a smaller, less noisy set of features, which can result in feature representation of reduced dimensionality without loss in performance.

Furthermore, wearable accelerometers can capture more subtle body movements in 3D space while not being affected by cross-contamination and occlusion and while maintaining the privacy of the conversation. Previous work has shown the utility of these signals in the detection of different actions occurring in a social setting, including, speaking, gesturing, and drinking [80, 81, 85, 128, 210]. We explore the addition of the acceleration modality for detecting speech activity.

Our contributions are the following:

- We collected an in-the-wild dataset with video and individual audio and body-worn accelerometer readings, in a crowded setting. We localized people in the scene via pose estimation. In contrast with previous work relying on visual annotations of speaking status [89], we obtained our ground truth automatically from high-quality voice recordings.
- We propose a method for speaking status detection that selects local features around pose keypoints, based on automatically-extracted body poses. We show through evaluation on our dataset that focusing on upper body keypoints, and head and hand keypoints in particular increases speaking status detection performance while decreasing the number of considered features, resulting in a faster and better-performing method. We reinforce our conclusions through evaluation on a second voice activity detection dataset with mild cross-contamination and a frontal view of subjects.

We show that performance improvements come from increased robustness against cross-contamination and analyze the complementarity between visual and accelera-

tion modalities. Through sensitivity analysis of hyper-parameters, we found that the dimensionality of the representation can be reduced five-fold (w.r.t. previous work) in the aggregation stage with a mild performance gain.

- We propose a method for multimodal prediction by late-fusing scores obtained from an acceleration stream, and analyze the situations in which modality information is redundant and complimentary.

3

3.2 RELATED WORK

3.2.1 VISUAL DETECTION OF ACTIONS

The action recognition field in computer vision has long studied the problem of recognizing human actions in videos. Traditional approaches include the extraction of dense trajectories [211], including their spatiotemporal descriptors, followed by an encoding method that transforms the trajectories from a video into a single high-dimensional video-level representation [212]. Fisher Vectors is the most prevalent and well-studied of such encoding methods due to its superior performance [213–215]. Improved dense trajectories [216] correct for camera motion and use bounding box human detections to filter out surrounding trajectories.

More recent approaches use the power of convolutional neural networks (CNNs). Two-stream networks feed static frames from the video into an appearance stream and optical flow computed between frames into a motion stream, the scores of which are fused for classification, effectively creating a separation between static and motion information [217]. Recently, 3D CNNs [218, 219] have gotten increased attention. Pre-training techniques using large datasets have demonstrated performance improvements over 2D CNNs and traditional approaches [220]. The more recent (2+1)D CNNs cover a middle ground by factorizing 3D convolutional filters into 2D spatial and 1D temporal convolutions [221], with however similar performance.

Although both 3D and (2+1)D CNNs deliver state-of-the-art results in most action recognition benchmarks, they are notorious for requiring training in large-scale datasets to achieve such performance. While using pre-trained models is possible for smaller datasets, the significant domain shifts between the kind of actions and viewpoints considered in large-scale datasets and a target dataset may make a model trained from scratch preferable. Despite the improvements in CNN efficiency, improved dense trajectories have featured close to state-of-the-art performance in the recent Charades dataset of everyday human activities [222, 223] while remaining competitive in traditional action recognition datasets among methods with no pre-training [220].

The development of faster and more accurate pose estimation and tracking systems capable of detecting the poses in a video at close to real-time [97, 98, 208, 209] has resulted in increased interest in pose-based action recognition systems. Previous work has assigned trajectories to joints, to later learn a separate bag-of-words per joint [126]. However, the method is designed and tested with frontal views of a single actor and therefore does not perform any filtering. While learning separate representation per joint works in such cases, it is impractical in a crowded setting where most joints are very frequently occluded.

Previous work [224] extracts RGB and optical flow patches around each pose keypoint and feeds them into one appearance and one motion CNN, effectively a two-stream network

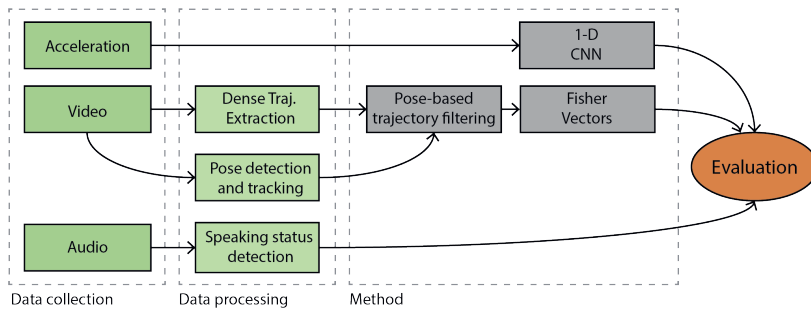


Figure 3.2: Overview of our work, from data collection to speaking status detection.

per keypoint. Similarly, Pishchulin experimented with multiple ways of combining ground truth pose and dense trajectory information, including filtering trajectories using square regions around pose keypoints [225]. Experiments were performed on a dataset with over 800 human activities. Performance improvements were highly dependent on the specific action being detected. No further details are given about the way trajectories were filtered.

3.2.2 NO-AUDIO SPEECH DETECTION

Despite the significant amount of work on generic action recognition and localization, interest in the problem of detecting the speaker in a conversation has been limited.

The most closely related vision works have addressed the problem of detecting speaking status from body movement information alone [226], introducing the dataset called Realvad for this purpose [227]. This dataset consists of a single-camera frontal recording of a panel discussion where 9 subjects take turns speaking. Even though this dataset does not contain a free conversation, it does capture a real event where the subject’s body movement is clearly visible. This dataset also has some cross-contamination, due to views of subjects overlapping with each other, albeit much less and less variable than in our free-standing conversation setting with multiple angles. The methods presented use domain adaptation to adapt CNN-extracted features of one speaker to another, using bounding boxes to localize and crop out the subject regions.

Related problems have also received some attention. One of them is the problem of speaker naming in movies [228, 229], where the goal is to localize and identify the speaking character in a movie or video. However, an important difference is that in movie naming the algorithm has access to the audio, and movie scenes tend to have clear frontal views of speakers, making it possible to rely on face detection and tracking [228].

On the other hand, acceleration readings have been used successfully for the assessment of human actions, more commonly for the recognition of daily activities like walking or running, but also for social actions including speech. Wang surveyed state-of-the-art deep learning approaches for sensor-based activity recognition, including accelerometers [230]. A study using chest-worn accelerometers established their ability to recognize actions like gesturing, laughing, and speaking in a free-standing social setting similar to ours [85]. The best results were found for the recognition of speaking, with more recent work improving on the methods used [81].

3.2.3 MULTIMODAL SPEECH DETECTION

Work on multimodal speaking gesture detection [80] is the most closely related to ours in method, presenting a multimodal method for gesture and speaking status detection in a crowded scene. The video-based method is inspired by Multiple Instance Learning for trajectory aggregation and classification and combined with an acceleration modality via late fusion. However, this study relies on human-annotated bounding boxes and speaking status annotations, and does not focus on solving the issue of cross-contamination.

Multimodal approaches in similar settings have found that a single accelerometer offers performance competitive with trajectory-based video recognition of speaking status [80, 128]. This work also found evidence that these modalities may complement each other and that a method will benefit from access to both.

3.2.4 SUMMARY

The main challenges in detecting speaking status from body movement in a crowded scene using existing methods are the subtlety (low intensity and visual saliency) of the movements involved (compared to the movements present in most action recognition datasets) and cross-contamination from other people in the scene. A smaller set of previous works looking specifically at speaking status detection without audio [80, 81, 226, 227] have addressed the first of these problems. However, to the best of our knowledge, the second problem remains unaddressed. In this paper, we attempt to understand and tackle this problem via existing action recognition methods using accelerometers and video data.

3.3 APPROACH

In this section, we justify and detail our decisions on the method using both the video and acceleration modalities. An overview of our approach, including data collection and processing is shown in figure 3.2.

3.3.1 VOICE ACTIVITY DETECTION FROM VIDEO

We start by describing the processing done on the video modality, including the process used to extract pose tracks from videos of human interaction.

POSE ESTIMATION

Accurate pose detection is central to our approach, which relies on it for local feature selection. Given the extent of previous works on pose estimation, we evaluated existing approaches in our scenario, which includes seldom-considered challenges including the crowdedness of the scene and the elevated angle of the camera. We considered two of the most well-known and maintained pose estimation methods: Openpose [97], a bottom-up approach based on part affinity fields; and AlphaPose, a top-down approach [98]. Both methods achieve competitive results in well-known pose estimation benchmarks like MPII Human Pose [231] and MS COCO Keypoints. Both methods delivered acceptable results in our videos, with some keypoints being more reliable than others. We decided to use Openpose (BODY25 model) due to its ability to consistently detect head and chest keypoints which we could use for person detection and tracking. Its real-time speed independent of the number of people in the frame was an extra advantage.

POSE TRACKING

Because most pose estimation algorithms, including OpenPose, work independently on individual frames, we needed a way to associate poses across frames. While this problem has been investigated in previous work, we found the existing PoseFlow [208] to be too computationally expensive for our use case, due to the large number of people in the scene. We therefore implemented a semi-automatic method to obtain tracks from individual frame detections. We opted for a computationally lighter method based on the observation that the chest keypoint was reliably detected and localized across frames.

Our method's objective is to create pose tracks by associating the chest keypoint across frames. Concretely, for each frame n of the video, the pose detector outputs a set of poses given by $Q_n = \{P_{n,1}, P_{n,2}, P_{n,M_n-1}, P_{n,M_n}\}$, where M_n is the number of people detected in frame n . A pose track J is a sequence of consecutive poses given by $J = \{P_{i_j}, \dots, P_{f_j}\}$, starting at frame i_j and ending at frame f_j .

With the same goal of a fast method, we decided on a step-wise approach, where the goal is to match poses in Q_n from frame n to tracks consisting of poses from all frames up to $n - 1$. Specifically, we solve the assignment problem between two sets of poses: Q_n and $\{P_{f_j} | J \in T_n \text{ and } n - f_j < R_{th}\}$, for $n = 1, 2, \dots, N$. This is repeated in order for $n = 1, \dots, N$. In other words, we compare poses in Q_n with the head of existing tracks whose last pose is not older than R_{th} frames, where R_{th} is an integer parameter.

The assignment problem is defined by a distance calculation. We defined the distance between two poses as the Euclidean distance between their chest keypoints $D(P_A, P_B) = \|\mathbf{P}_A^{chest} - \mathbf{P}_B^{chest}\|$. We solve the assignment problem via the Hungarian algorithm. We add a maximum distance threshold D_{th} for assignment, such that if $D(P_A, P_B) > D_{th}$, then P_A and P_B cannot be assigned to each other. Assigned keypoints are added to the corresponding track and unassigned keypoints are assigned to a new track. When a new pose in frame n is matched to a pose in a frame between $n - 30, \dots, n - 2$ (ie. a pose not in frame $n - 1$), we imputed the keypoints via linear interpolation to maintain the continuity of the track.

Parameters $R_{th} = 50$ and $D_{th} = 30$ frames were set based on experiments on a subset of the tracks. This approach resulted in high-quality tracks, with only a few person switches due to subjects walking in front of one another. We found the one-second threshold to work well in eliminating issues of consistency across frames without introducing significant errors.

Because our goal was to obtain high-quality tracks to reliably test our recognition method (see next sections), we manually inspected the dataset for track switches and corrected them by splitting the tracks. We further assigned tracks to the corresponding ID of the participant to be able to associate with the personal acceleration readings.

DENSE TRAJECTORIES

Research in psychology and social signal processing has shown that body gestures, especially hand gestures, are closely synchronized with speech [156, 232]. Because of the difficulty of accessing facial information in our setting, our method aims to capture primarily such gestures and overall body movement. For our video-based detection method, we relied on dense trajectories due to their ability to track such salient movements. Additionally, the relatively small size and non-standard viewpoint of our dataset make a more simple method trained from scratch preferable to a method based on pre-trained CNNs. Given the

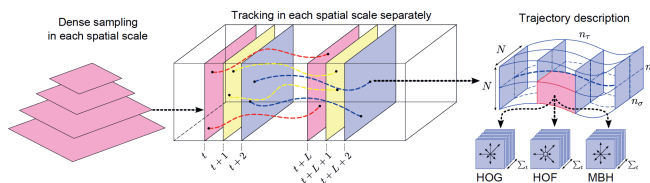


Figure 3.3: Dense trajectories are created by sampling interest points from multiple scales. These are tracked by following the optical flow fields over L frames. Finally, features are extracted to describe the volume around the trajectory. [211]

3

absence of camera motion in our videos, we used the original dense trajectories [211], and not improved trajectories [216].

Dense trajectories [211] were proposed in action recognition literature for the classification of short videos labeled with the action being performed in them. Dense trajectories are extracted by sampling feature points and tracking them for the following frames using optical flow. Feature points are sampled on a grid spaced by $W = 5$ pixels in 8 spatial scales spaced by a factor of $1/\sqrt{2}$. Points are then tracked using a dense optical flow field. This is done for $L = 15$ frames to avoid drift from longer trajectories. The spatio-temporal volume in a neighborhood of size $N = 32$ pixels around the trajectory is then described using histograms of gradients (HOG), histograms of optical flow (HOF) and motion boundary histograms (MBH) features extracted from cells dividing the volume in a grid of size $n_l \times n_l \times n_t = 2 \times 2 \times 3$ (see figure 3.3). While HOG features are predominantly visual, HOF features capture more temporal information. Motion boundary histograms capture both visual and temporal information.

The mentioned dense trajectory parameters: length L , step size W , neighborhood size n_l , n_t were set to their default values, which are replicated in most previous action recognition literature using trajectories and shown to be close to optimal on different datasets [211]. The final dense trajectory descriptor is the concatenation of the trajectory (size 30), HOG (size 96), HOF (size 108), and MBH (size 192) vectors for a total of $D = 426$ dimensions describing a video segment.

POSE-BASED FILTERING OF DENSE TRAJECTORIES

The goal of our filtering method is to reduce the effect of spurious trajectories due to cross-contamination by people other than the target subject. At the same time, we do not attempt to precisely segment the subject. While precise segmentations would be ideal, person instance segmentation methods [99] add significant computational complexity and in preliminary experiments had worse performance in detecting joints in our dataset.

First, due to the frequent occlusion of lower-body keypoints as a result of the crowd density and viewing angle, we only consider 8 upper-body keypoints (see figure 3.4). We obtain a single head keypoint by averaging the eyes, nose, and ears keypoints output by the pose estimator. We pick trajectories starting within a radius of these 8 keypoints, and filter out the rest. Because the estimated trajectories accumulate errors [211], their initial position (first point) is the most reliable in terms of matching the position of a possibly salient point.

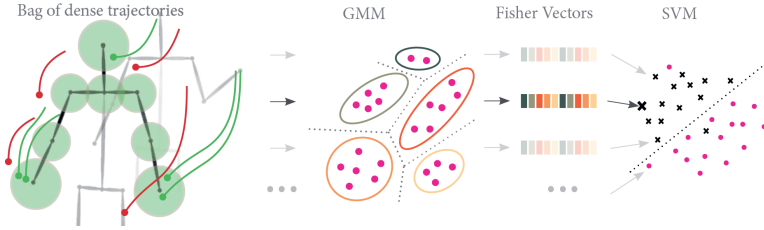


Figure 3.4: Our video-based approach selects trajectories around pose keypoints.

We filter such that trajectory x originating at $\mathbf{o}_x = (o_x^{\hat{x}}, o_x^{\hat{y}})$ starting at frame n is compared to the target's pose joint keypoints of the same frame $P_n = \{\mathbf{p}_{n,1}, \dots, \mathbf{p}_{n,j}, \dots, \mathbf{p}_{n,8}\}; \mathbf{p}_{n,j} = (p_{n,j}^{\hat{x}}, p_{n,j}^{\hat{y}})$ of frame n . Trajectory x is selected if $\|\mathbf{o}_x, \mathbf{p}_{n,j}\| < R_{n,j}$ for any joint j . We also compare x with poses in P_{n-1} and P_{n+1} using the same criterion. We found comparing with 3 frames to add significant robustness against inconsistent keypoint detections (common due to the frame-wise pose estimation). Using larger comparison windows tended to add increasingly more noise.

In our data, participants vary greatly in pixel size depending on their distance relative to the camera. To account for this, we scale $R_{n,j}$ according to the distance from the camera of the particular keypoint, ie. $R_{n,j} = R_j * S(p_{n,j})$, where S computes a scaling factor that depends on the position of the point and R_j are hyperparameters.

We obtain the scaling factor S via the camera-to-ground plane homography. We compute the transformation A between the ground plane and the image plane using marks placed at known distances on the floor plane during data collection. This allows us to approximate a scale ratio between point $p_{n,j}$ and a reference point p_r in the image plane as $\|A(p_{n,j}), A(p_{n,j} + \Delta p)\| / \|A(p_r), A(p_r + \Delta p)\|$ where Δp is an arbitrarily small displacement vector.

EXTRACTING FISHER VECTOR REPRESENTATION

The chosen trajectories from different joints must be aggregated into a video-level representation. We use Fisher vectors [215], the state-of-the-art method for dense trajectory aggregation. Fisher vectors, and in particular their improved variant [214] have been found to perform remarkably well in large-scale datasets of daily activities like the recent Charades [222, 233].

Fisher vectors provide a compact feature representation from an arbitrary number of dense trajectories. Let $X = \{\mathbf{x}_t, t = 1 \dots T\}$ be the set of T dense trajectories of dimensionality $D = 426$ selected from keypoint regions and u_{λ} be the probability density function with parameters λ . The Fisher score is defined as the gradient of the log-likelihood over X , with respect to the model parameters:

$$\mathbf{G}_{\lambda}^X = \frac{1}{T} \nabla_{\lambda} \log u_{\lambda}(X) \quad (3.1)$$

The Fisher vector is a normalized version of the Fisher score:

$$\mathcal{G}_{\lambda}^X = \mathbf{L}_{\lambda} \mathbf{G}_{\lambda}^X \quad (3.2)$$

where normalization by L_λ corresponds to the whitening of the dimensions, where a generative model can take the place of u_λ . Normally a Gaussian mixture model (GMM) with K components and diagonal covariance matrices is used. The parameters λ of a GMM are $\lambda = \{w_i, \mu_i, \sigma_i^2, i = 1, \dots, K\}$, where w_i , μ_i and σ_i^2 are the mixture weight, mean vector and diagonal of the covariance matrix of Gaussian i . However, only mean and standard deviation are used because mixture weights add little additional information [214]. Under the assumption of independence of local descriptors:

$$\mathbf{G}_\lambda^{\mathbf{X}} = \frac{1}{T} \sum_{t=1}^T \nabla_\lambda \log u_\lambda(\mathbf{x}_t) \quad (3.3)$$

Let $\gamma_t(i)$ be the soft assignment of descriptor \mathbf{x}_t to Gaussian i :

$$\gamma_t(i) = \frac{w_i u_i(\mathbf{x}_t)}{\sum_{j=1}^K w_j u_j(\mathbf{x}_t)} \quad (3.4)$$

Calculation of the gradients leads to:

$$\mathbf{G}_{\mu,i}^{\mathbf{X}} = \frac{1}{T} \frac{1}{\sqrt{w_i}} \sum_{t=1}^T \gamma_t(i) \left(\frac{\mathbf{x}_t - \mu_i}{\sigma_i} \right) \quad (3.5)$$

$$\mathbf{G}_{\sigma,i}^{\mathbf{X}} = \frac{1}{T} \frac{1}{\sqrt{2w_i}} \sum_{t=1}^T \gamma_t(i) \left[\frac{(\mathbf{x}_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \quad (3.6)$$

where the division between vectors is term-by-term. The Fisher Vector aggregates all gradients into a vector of $2KD$ dimensions. For $K = 256$ (used in previous work [214]), this is a 218112-dimensional vector. Finally, Fisher vectors are normalized by dividing by their L2 norm and then power-normalized with $f(z) = \text{sign}(z) \sqrt{|z|}$. These techniques have been shown to increase the ability of the Fisher Vector to detect subtle elements, and reduce the sparsity of Fisher vectors [214] respectively.

A kernel on these gradients is defined as:

$$K(X, Y) = G_\lambda^{X'} F_\lambda^{-1} G_\lambda^Y = G_\lambda^{X'} L_\lambda'^{-1} L_\lambda^{-1} G_\lambda^Y \quad (3.7)$$

where F_λ is symmetric and positive definite, and generally approximated such that normalization by L_λ corresponds to a simple whitening of the dimensions.

Linear methods (traditionally linear SVM) are standard for the classification of the FVs [30,38] because learning a linear classifier on the FVs is equivalent to learning a classifier using the Fisher kernel (kernel trick) and linear methods have delivered good results in previous work [33].

3.3.2 VOICE ACTIVITY ESTIMATION FROM WEARABLE ACCELERATION

Due to the good results obtained by deep methods in sensor-based activity recognition [230], we use a one-dimensional CNN for acceleration-based detection. Just as previous work which makes use of relatively shallow CNNs for detecting actions from a single-accelerometer [234, 235] we use a flattened version of the two-dimensional AlexNet [236], where we preserve the ratios between the number of channels. Figure 3.5 shows the

architecture. Input data has 3 channels corresponding to axes X, Y, and Z of the acceleration signal. Filter sizes are 5 for the first convolutional layer and 3 for the other layers, with unit padding. As with AlexNet, the first, second, and last layers are followed by a max-pooling layer with kernel size 3 and stride of 2.

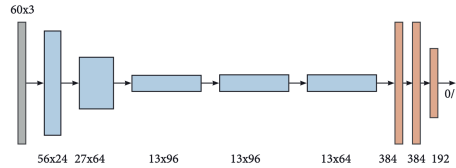


Figure 3.5: Architecture of the 1D-CNN used.

The input to the CNN is pre-processed by subtracting the mean and dividing by the standard deviation, for each axis, to reduce the effect of gravity and device miscalibration.

3.3.3 MULTIMODAL FUSION

The unimodal classifiers described above provide a posterior probability of the voice activity. We combine both modalities via late fusion to obtain multimodal scores. We opt for this approach because we expect acceleration to be complimentary to video in many cases when gestures or other vocalization-associated body movements are hard to observe due to occlusion or orientation. While the acceleration signal encodes chest motion specifically, Fisher Vectors encodes a mixture of visual and motion information from different body parts.

3.4 DATASETS

We used two datasets with available speaking status annotations to validate our approach. The two datasets differ in having significantly distinct views of the subjects and in the setting in which they were collected. Both datasets, however, share the issue of cross-contamination in the video modality. The first dataset has an elevated side-view of subjects and was collected by us by recording an in-the-wild mingling event, which also included accelerometer sensors. This dataset has free interaction between 43 recorded participants, and therefore frequent turn-taking. The second dataset, Realvad was published as part of a paper analyzing no-audio speaking status detection [227]. It was collected during a panel session where participants took turns speaking, and has therefore less frequent turn changes.

3.4.1 FREE STANDING DATASET

Detection of speaking status in the wild requires the collection of a dataset in which social interaction occurs with as little intervention as possible. To this end, our dataset was collected during a special event organized by a business networking group. Most participants in the event meet regularly and many but not all of them knew each other. Participants were informed beforehand that this particular meeting would be recorded. As they arrived to the event, they were asked for consent after being further informed about

the data collection. They were allowed to choose which sensors to wear (microphone, accelerometer, or both) or to not participate in the data collection. They were informed about a clearly-delimited video zone where they would be recorded by our video cameras. Of about 100 attendees, 43 consented to wearing sensors. This process was approved by the ethics board of the university beforehand.

During the event, most of the interaction consisted of free-standing conversation. Participants were free to move around and talk as they pleased, and they were video-recorded when inside the video zone. Because most participants were acquainted with each other and this was a special event commemorating an anniversary of their organization, conversations were mostly friendly and sporadic.

PARTICIPANTS.

43 participants took part in the data collection. Of them, 20 were male and 23 female.

SENSOR SETUP.

We collected data using the following sensors:

- A custom-made wearable accelerometer sensor hung around the neck and rested on the chest like a smart ID badge.
- Lavalier microphones attached to the face to record speaking activity. Microphones were attached to a Sennheiser SK2000 transmitter. Audio was recorded at 48kHz.
- 12 overhead cameras and four side-elevated cameras were placed above and in the corners of a video zone. In this work, we only make use of the side elevated cameras.

Because many participants chose to only wear one of the sensors or to not enter the video zone, and because of the malfunction of some of our wearable devices, not all modalities were available for all participants. Table 3.1 shows the dataset statistics, where 17 out of a possible 43 participants had data from both modalities available to them. This highlights the challenges with capturing in-the-wild data where participants can choose what data to provide. However, we can be more confident of the realism of the data compared to more controlled settings.

DATA ANALYSIS

The main goal of this work is to investigate the feasibility of a multimodal approach for speaking status detection in crowded settings, where the video stream is based on pose estimation and the wearable acceleration stream is a complementary modality. In this section, we detail the data preparation process, including obtaining speaking status annotations, detecting and tracking poses, and creating the dataset of speaking and non-speaking examples.

Our recordings included periods in which the participants were expected to attend to a speaker or listen to a performance. We used the video modality to manually find and eliminate such segments, as they deviate from our setting of interest.

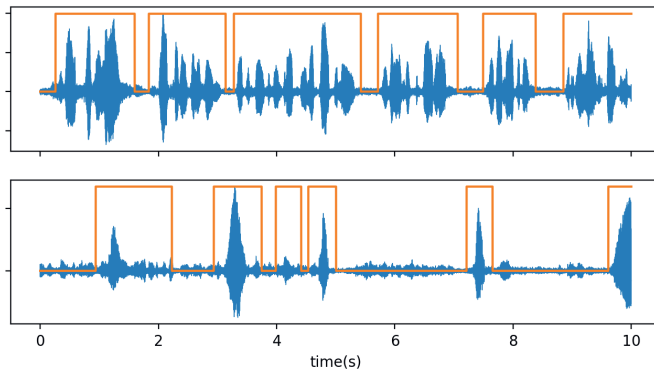


Figure 3.6: Example of our speaking status detections showing the audio waveform and VAD binary predictions for two data segments. In the first case, the speaker takes few pauses, in the second it is back-channeling to its interlocutor.

Table 3.1: Statistics of our speaking status dataset. The number of positive examples appears in parentheses.

Modalities	Participants	Num. Examples	Hours
Video	24	18639 (10635)	15.53
Video & Acceleration	17	12309 (7695)	10.26

3.4.2 AUTOMATIC SPEAKING STATUS ANNOTATION

Due to the availability of high-quality audio recordings of our subjects, we first explored the automatic annotation of speaking activity via voice activity detection (VAD) algorithms. Due to the closeness to the mouth of our head-worn microphones, there was a significant difference in energy between the speech of the speaker and background speech. The presence of background noise, however, poses a challenge for VAD.

We first investigated the feasibility of using pre-trained neural models, trained on both the AMI dataset, and the more diverse and challenging DIHARD dataset [237], through the *pyannote.audio* package [238, 239]. However, we found these methods to be too sensitive for our use case, detecting most background speech with high confidence.

Therefore, we relied on the rVAD method for robust voice activity detection [240]. rVAD relies on pitch detection, applies several de-noising passes, and directly accounts for signal energy differences in segmentation of the speech signal. We used the full version of the rVAD detector to produce binary speaking status outputs for our participant recordings at 100Hz. We found this method to reliably segment speaker voice activity from background noises. Figure 3.6 shows an example of two output segmentations.

3.4.3 REALVAD DATASET

We also test on the Realvad dataset presented in [227]. Although this dataset has significantly less cross-contamination, we are also interested in testing our approach on a dataset where body parts are more consistently visible, to better understand their relative importance as indicators of speaking. Furthermore, frontal shots like those in Realvad are

common, and we wish to understand how the performance of the method changes under such condition.

3.4.4 DATA PRE-PROCESSING

For our experiments, we split the obtained tracks and accelerometer readings in 3-second segments, each constituting one data sample. We did so because previous work has found windows of 3s to be maximally informative in speaking status detection tasks [81]. We labeled examples using a threshold on the fraction of positive VAD labels in the segment. Rather than using majority voting (0.5 threshold), we opted for a more aggressive threshold of 0.25 that would label most examples with speech activity as positives, which resulted in a more balanced dataset. Table 3.1 shows the label statistics.

3

3.5 EXPERIMENTS

We extract dense trajectories with a length of $L = 15$ and a sampling stride of $W = 5$. These settings were found to be optimal in the original paper on dense trajectories [211] and have been used as standard in more recent work [212, 216].

We train the GMM using a sample of 100000 trajectories, to which we apply Principal Component Analysis (PCA) preserving 95% of the variance and whitening. We set $K = 256$ components for the GMM and apply power normalization ($\alpha = 0.5$) and L2 normalization to the Fisher vectors. All of these parameters and transformations were found to be optimal across a variety of datasets in previous work on best practices for training Fisher vector models [214]. We perform classification using a linear SVM with an L2 regularizer, following the same literature [214]. Training was done via stochastic gradient descent (SGD). The optimal regularization parameter was found via cross-validation.

We tuned the hyperparameters R_j of our pose-aligned trajectories approach via experiments on a small subset of data. Intuitively, these parameters determine the radius around each body keypoint from which trajectories are sampled. For simplicity, we considered $R_j = R\forall j$ (ie. the same radius is considered around every keypoint) and tested a set of 5 parameter settings (which we determined visually from data samples) on the held-out set via 4-fold cross-validation, which led us to a setting for R .

For the acceleration stream, we train the network using a binary cross-entropy loss and the Adam optimizer. Late fusion is performed by training a logistic regressor without regularization on the output scores of both modalities.

We evaluate all models via 10-fold cross-validation. How the data is split is relevant in our case. A person-level cross-validation split would be ideal to avoid significant dependencies between examples in the training and test sets. Due to the low number of participants, however, we opted for a split where every person-camera combination is considered one group, such that examples of the same person viewed from the same camera are always put together in either the train or the test set.

We use the area under the ROC curve (AUC) as the main evaluation metric as it quantifies the ability of the method to separate positive and negative labels in the output space while being robust against dataset imbalance. We use Platt scaling to obtain probability scores from the SVM outputs.

Table 3.2: Results of 10-fold cross-validated experiments comparing our pose-based selection method with other methods based on dense trajectories. FV stands for Fisher Vectors.

Method	Trajectories / example	AUC
FV - Full	1550.18 (1069.01)	0.686 (0.015)
FV - Sampled	527.71 (345.93)	0.678 (0.019)
FV - UpperBody	619.67 (519.31)	0.713 (0.016)
FV - HandsAndHead	522.98 (458.82)	0.715 (0.020)

3.5.1 POSE-BASED FILTERING

We start by analyzing the effect that pose-aligned trajectory selection has on speaking status detection. Previous work [80] has shown that hands and arms produce informative trajectories for speaking status detection. We hypothesized that selecting trajectories around skeleton keypoints will result in a more informative, less noisy set of trajectories that will achieve greater generalization.

To test the improvement obtained from trajectory selection, we compared our method with a traditional bounding box approach. A sample’s bounding box was obtained from the pose tracks by computing the smallest box that contains the subject’s skeleton. For consistency, we computed bounding boxes for the upper body only, given the frequent occlusion of the lower body. A padding scaled using A (see section 3.3.1) was added such that the subjects’ upper body would be contained in the bounding box.

Table 3.2 shows the results and the number of trajectories used by each method. It indicates that our method improves over the baseline despite using only a subset of trajectories. We test two variants of our method. In the first one (*FV - UpperBody*) we extract trajectories around all upper body keypoints as presented in section 3.3.1. In the second (*FV - HandsAndHead*) we select trajectories only around the head keypoint, and both wrist keypoints. Interestingly, this last method delivers the best results, while using only 34% of the original trajectories on average. Both methods improve over the baseline (*FV-Full*) by a significant margin. As an extra reference, we add a *FV-Sampled* method which corresponds to random sampling of the trajectories with probability $p = 0.34$, such that each example has on average the same number as selected by our method.

ROLE OF CROSS-CONTAMINATION

Since the initial motivation of our method is to avoid cross-contamination in a crowded setting, we further investigated its role in our results. To this end, we gave every example a cross-contamination score. First, for a target bounding box B in frame n , we compute its cross-contamination score as:

$$CC_B^n = \frac{\sum_i \text{Intersection}(B, B_i^n)}{\text{Area}(B)}$$

where B_i^n is the bounding box of pose i in frame n . The score is the ratio between the intersection of the target pose’s bounding box with all other detected poses’ bounding boxes and the area of the target bounding box. To obtain an example-level score we took the median of its frame scores to remove the effect of outliers due to differences in the estimated poses.

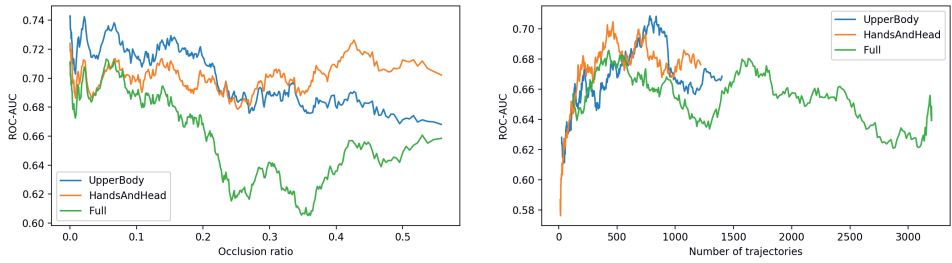


Figure 3.7: Left: AUC scores as a function of the cross-contamination score. Right: change in AUC scores with the number of trajectories in the example.



Figure 3.8: Some examples of cross-contamination.

While this is not a perfect measure of cross-contamination, in our experiments we found this measure to consistently assign high scores to segments in which the target was significantly occluded by another person, and decrease with less severe cases of cross-contamination.

Figure 3.7 shows the results of plotting the ROC AUC score as a function of the cross-contamination score. Here, it can be seen that our method is especially stable regardless of cross-contamination level, while a bounding box is inevitably affected by it. Interestingly, the model considering only hands and head is significantly more robust than the one considering all upper body keypoints. The number of trajectories in the example is also shown to have an influence, with examples with few trajectories being more frequently misclassified.

Figure 3.8 shows some examples of segments with high cross-contamination scores where the target subject's interlocutor occludes the subject's body. Points indicate the origin of a trajectory, with green points indicating trajectories selected by our method and white ones being discarded. In the third case, our method avoids significant contamination from the target's interlocutor due to occlusion.

ROLE OF BODY PARTS AND DESCRIPTORS

In this section, we evaluate the relative importance of different body parts in the results obtained, using our dataset and the Realvad dataset. Because of the results of the previous section, indicating that the method works equally well with only hand and head keypoints, we evaluate only on these keypoints.

Table 3.3: AUC results of 10-fold cross-validated experiments comparing the effect of using only information from hand or head keypoints. *excl.* indicates that examples without the corresponding body part have been excluded from the computation.

Dataset	Body part	Descriptors				
		Trajectory	HOG	HOF	MBH	All
FreeStanding	Head	.6751	.6958	.6556	.7018	.7438
	Head (<i>excl.</i>)	.6407	.7056	.6580	.7138	.7467
	Hands	.5835	.5927	.5974	.6123	.5865
	Hands (<i>excl.</i>)	.5620	.5288	.5755	.5842	.6015
	Hands&Head	.6834	.6965	.6680	.7163	.7372
	Hands&Head (<i>excl.</i>)	.6648	.7138	.6717	.7177	.7484
Realvad	Head	.7980	.7464	.8115	.8521	.8017
	Hands	.8230	.7681	.8334	.8332	.8475
	Hands&Head	.8650	.7876	.8537	.8970	.8743

Table 3.3 shows the results of an ablation study where we remove features from either the hands or the head. We also reduce the descriptor set by only taking HOG, HOF or MBH descriptors for all trajectories. This effectively reduces the dimensionality of the Fisher vectors. We are most interested in the method’s performance when all features are used. The method using only head trajectories had in some cases better performance. The differences between head-only and head-and-hands methods are however statistically insignificant. The results did show that the hand movement information is less important, with a nearly 10% difference in most cases.

The situation differs for the Realvad dataset, where both hands and head are similarly informative. The higher scores compared to the freestanding dataset can be explained by the setting, where people are sitting and relatively close to the camera. Interestingly, the classifier with only MBH features performed better than when all features were used, possibly due to the influence of MBH features from the head. Inspection of the data revealed a possible reason for hands and head having similar influence: in Realvad listeners tended to keep their hands on their laps (or otherwise completely still) most of the time, while the current speaker would frequently make hand gestures. Overall, however, it stands out that head movements are a strong predictor across both datasets.

Therefore, we conclude that the success of the filtering approach is driven mainly by information from head trajectories. This can be explained by the fact that in our freestanding dataset hands are more frequently occluded, especially behind the body depending on the person’s orientation. To understand how occlusion of the hands affects the results, we also computed AUC scores excluding examples where the hands and/or head are not visible. These results are indicated with *excl.* in table 3.3 and they show that the performance improvement is mild even when considering only examples where the body part is visible. This strongly suggests that the head area contains most of the information relevant to speaking status detection.

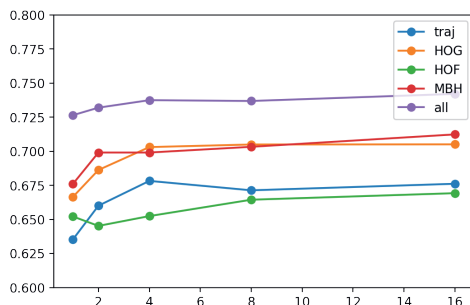


Figure 3.9: AUC scores using different GMM sizes.

Table 3.4: Results of 10-fold cross-validated experiments comparing our 1-D CNN acceleration-based method with previous work.

Method	AUC
PSD + Logistic Regression	0.698 (0.025)
1D-CNN	0.738 (0.029)

SENSITIVITY ANALYSIS OF PARAMETERS

We perform a sensitivity analysis to understand the importance of different hyperparameters in model performance. We keep the dense trajectory hyperparameters fixed following the standards of previous literature, which have found them to be close to optimal for different datasets. We focus on the filtering and Fisher vector hyperparameters. We start by experimenting with the GMM size K .

Figure 3.9 shows the AUC scores obtained for different numbers of components in the GMM. We show GMM sizes between 4 and 16 because this is where we could observe the most variation. GMM sizes greater than 16 resulted in no significant increase in performance, indicating that the method can be simplified further by using a GMM smaller than the standard 256. This indicates that features likely follow a distribution with few modes. This can be the case in our dataset due to its uniform setting, in contrast with the diversity of settings, backgrounds, subjects, and even video qualities present in action recognition datasets.

3.5.2 MULTIMODAL SPEAKING STATUS DETECTION

Table 3.4 shows the results of our speaking status detection method compared with previous work. We compare with a baseline consisting of augmenting the acceleration signal with the magnitude and absolute value of each axis, followed by the computation of a power spectral density (PSD), binned into 8 logarithmically spaced bins. We follow the implementation detailed in previous work, including classification using a logistic regressor [81]. We did not compare with the personalized models proposed in such work due to the very low number of participants in our dataset. The logistic regressor hyperparameter C was tuned in a nested cross-validation loop.

Table 3.5: Results of 10-fold cross-validated experiments on our multimodal fusion approach.

Method	AUC
FV - HandsAndHead	0.720 (0.031)
1-D CNN	0.738 (0.029)
Multimodal	0.763 (0.027)

The results in table 3.4 indicate that our 1-D CNN outperforms previous work. This is no surprise given that a CNN can learn complex features directly from the signal.

Table 3.5 shows the results of our multimodal approach with late fusion. These experiments are run only on the subset of the examples with both video and accelerometer data. Results suggest that both modalities indeed have a high degree of complementarity.

3.6 LIMITATIONS AND FUTURE WORK

There are several limitations of our work that we consider important to discuss.

First and foremost, although we showed that automatic pose estimation methods are viable for person localization, it is also true that the extracted poses are not always reliable. The camera angle, illumination, and occlusion can significantly affect the quality of the pose estimation step. Our approach is not robust against mistakes by the pose estimator. This can potentially be improved through trajectory weighting instead of selection. In weighting, trajectories close to skeleton keypoints are weighted more than trajectories far from the keypoints. In this way, when few of the visible body keypoints are found by the pose estimator, the method falls back to considering most trajectories with similar importance. However, this method has the disadvantage of introducing some noise trajectories which could potentially offset its benefits. We are similarly interested in the utility of person instance segmentation methods in this step, but improvements in the quality and speed of such methods are necessary.

Second, our method requires the extraction of dense trajectories. While dense trajectories are still an excellent option due to their speed, performance, and easy parallelization of both extraction and processing with methods like Fisher Vectors, they have some clear drawbacks against neural-based approaches like their large and variable space requirements, and the inability to learn their parameters. Although not the primary goal of our work, we observed that one drawback of dense trajectories in this task in particular comes from the filtering of trajectories that are too static. This is necessary with dense trajectories to prevent an explosion in the number of trajectories but means that subtle cues related to speaking, like slight head or mouth movements, might be filtered out, leaving the method with no information in cases when there is no long-range movement of the limbs or body. For these reasons, we are interested in future work which combines these ideas with neural approaches, in an attempt to understand the importance of such subtle cues.

Regarding the acceleration modality, due to the low dimensionality of the input, we think there is little benefit to be obtained from larger models. We believe future research on using this modality for speaking status detection should explore using more and higher frequency sensors. Research suggests that wrist and chest-worn sensors can be informative

of speaking status while remaining relatively unobtrusive and privacy-preserving. Regarding modality fusion, we are interested in exploring smarter fusion approaches based on the observation that acceleration should have more influence in the prediction when the information available to the video stream is low.

3.7 CONCLUSIONS

In this work, we presented a multimodal method combining wearable sensors with a pose-based video approach for speaking status detection in crowded settings, where the video modality can be severely affected by occlusion and cross-contamination.

Using a dataset collected in the wild and annotated automatically for speaking status using high-quality voice recordings, we showed that pose detections from a state-of-the-art method are not only viable for person detection and tracking in a crowded scene, but that leveraging pose information in the action recognition stage improved performance while at the same time reducing the number of local features considered by the action recognition stage. This indicates a less noisy, more informative representation. The analysis of our method revealed that it is in cases of occlusion that our method improves over the holistic approach, underscoring the advantage of using poses for person localization in a crowded scene.

Finally, the significantly improved performance of the multimodal approach indicated that the video and acceleration modalities were complimentary.

We hope these results will help inspire and adapt similar approaches for improving the quality and speed with which machines can understand a crowded scene while reducing human and computational time expenses.

ACKNOWLEDGEMENTS

This research is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

4

CONFLAB: A DATA COLLECTION CONCEPT, DATASET, AND BENCHMARK FOR MACHINE ANALYSIS OF FREE-STANDING SOCIAL INTERACTIONS IN THE WILD

4

Recording the dynamics of unscripted human interactions in the wild is challenging due to the delicate trade-offs between several factors: participant privacy, ecological validity, data fidelity, and logistical overheads. To address these, following a datasets for the community by the community ethos, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of in-the-wild free-standing social conversations. For the first instantiation of ConfLab described here, we organized a real-life professional networking event at a major international conference. Involving 48 conference attendees, the dataset captures a diverse mix of status, acquaintance, and networking motivations. Our capture setup improves upon the data fidelity of prior in-the-wild datasets while retaining privacy sensitivity: 8 videos (1920 × 1080, 60 fps) from a non-invasive overhead view, and custom wearable sensors with onboard recording of body motion (full 9-axis IMU), privacy-preserving low-frequency audio (1250 Hz), and Bluetooth-based proximity. Additionally, we developed custom solutions for distributed hardware synchronization at acquisition and time-efficient continuous annotation of body keypoints and actions at high sampling rates. Our benchmarks showcase some of the open research tasks related to in-the-wild privacy-preserving social data analysis: keypoints detection from overhead camera views, skeleton-based no-audio speaker detection, and F-formation detection.



Figure 4.1: Snapshot of the interaction area from our cameras. We annotated only cameras highlighted with red borders (high scene overlap). For a clearer visual impression of the scene, we omit cameras 1 (few people recorded) and 5 (failed early in the event). Faces blurred to preserve privacy.

4.1 INTRODUCTION

4

A crucial challenge towards developing artificial socially intelligent systems is understanding how *real-life* situational contexts affect social human behavior [241]. Social science findings indeed show that the dynamics of how we conduct daily interactions vary significantly depending on the social situation [242–244]. Unfortunately, such dynamics are not adequately captured by many data collection setups where role-played or scripted scenarios are typical [245].

In this paper, we address the problem of collecting a privacy-sensitive dataset of unscripted social dynamics of real-life relationships where encounters can influence someone’s daily life. We argue that doing so requires recording these exchanges in the natural ecology, requiring an approach different from the typical setup of locally organized studies. Specifically, we focus on free-standing interactions within the setting of an international conference (see Figure 4.1).

Recording an international community in its natural habitat is characterized by several intersecting challenges: an intrinsic trade-off exists between data fidelity, ecological validity, and privacy preservation. For ecological validity, a non-invasive capture setup is essential for mitigating any influence on behavior naturalness [247–249]. The most common solution involves mounting cameras from aerial perspectives such as top-down [89, 90] and elevated-side views [88, 94, 250]. Now elevated-side views make it easy to capture sensitive personal information such as faces, which leads to several ethical concerns. For instance, capturing faces has been related to harmful downstream surveillance applications [251]. Besides, state-of-the-art (SOTA) body-keypoint estimation techniques perform poorly on aerial perspectives [89, 252], making the extraction of automatic pose annotations challenging

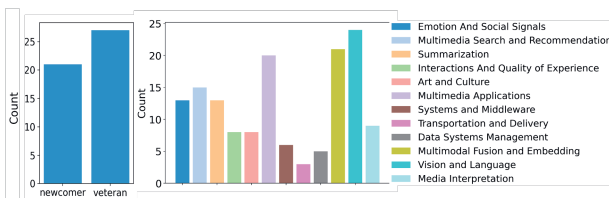


Figure 4.2: Frequency of newcomer/veteran participants (left) and reported research interests (right).



Figure 4.3: Keypoint detection using pre-trained RSN [246]. Additional SOTA results are in Appendix 4.F.1

(Figure 4.3). To avoid such issues, some researchers have turned to more privacy-preserving wearable sensors shown to benefit many behavior analysis tasks [81, 198, 249].

In all, the closest related datasets (see Table 4.1) suffer from several technical limitations precluding the analysis and modeling of fine-grained social behavior: (i) lack of articulated pose annotations; (ii) a limited number of people in the scene, preventing complex interactions such as group splitting/merging behaviors, and (iii) an inadequate data sampling-rate and synchronization-latency to study time-sensitive social phenomena [253, Sec. 3.3].

To address all these limitations, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of ecologically valid social settings. From the first instantiation of ConfLab, we provide a high-fidelity dataset of 48 participants at a professional networking event.

Methodological Contributions: We describe a data collection design that captures a diverse mix of real levels of seniority, acquaintance, affiliation, and motivation to network (see Figure 4.2). This was achieved by organizing ConfLab as part of a major international scientific conference. ConfLab had these goals: (i) a data collection effort following a *by the community for the community* ethos: the more volunteers, the more data, (ii) volunteers who potentially use the data can experience first-hand potential privacy and ethical considerations related to sharing their own data, (iii) in light of recent data sourcing issues [251, 254], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

Technical Contributions: (i) **aerial-view articulated pose:** our annotations of 17 full-body keypoints enable improvements in (a) pose estimation and tracking, (b) pose-based recognition of social actions (under-explored in the top-down perspective), (c) pose-based F-formation estimation (has not been possible from prior work [90, 91, 255, 256]), and (d) the direct study of interaction dynamics using full body poses (previously limited to lab settings [257]). (ii) **subtle body dynamics:** we are the first to use a full 9-axis Inertial Measurement Unit (IMU) enabling a richer representation of behavior at higher sample rates; previous rates were found to be insufficient for downstream tasks [81]. (iii) **enabling finer temporal-scale research questions:** a sub-second crossmodal latency of ~ 13 ms along with higher sampling rate of features (60 fps video, 56 Hz IMU) opens the gateway for the in-the-wild study of nuanced time-sensitive social behaviors like mimicry and synchrony.

4.2 RELATED WORK

Early datasets of in-the-wild social events either spanned only a few minutes (e.g. Coffee Break [94]), or were recorded at such a large distance from the participants that performing robust, automated person detection or tracking with SOTA approaches was non-trivial (e.g. Idiap Poster Data [90]). More recently, two different strategies have emerged to circumvent such issues.

One approach involves fully instrumented labs with a high-resolution multi-camera setup for video and audio data. Here automatic detectors [257, 259, 260] could be applied to obtain poses. This circumvents the cost- and labor-intensive process of manually labeling

Table 4.1: Comparison of Conflab with prior datasets of free-standing conversation groups in in-the-wild social interaction settings. N: number of people in the recorded scene. Conflab is the first and only social interaction dataset that offers skeletal keypoints and speaking status at high annotation resolution, as well as hardware synchronized camera and multimodal wearable signals at high resolution.

Dataset	N	Video	Manual Annotations	Wearable Signals	Synchronization
Cocktail [250]†	7	512 × 384	F-formations (20 and 30 min, 1/5 Hz)	None	Unknown
CoffeeBreak [94]	14	1440 × 1080	F-formations (130 frames in two sequences)	None	None
IDIAP [90]	> 50	180 min; 654 × 439 20 fps	F-formations (82 independent frames)	None	None
SALSA [88]†	18	60 min; 1024 × 768 15 fps	Bounding boxes (30 min) Head & body ori. (30 min) F-formations (60 min) (all 1/3 Hz)	Audio MFCCs (30 Hz) Acceleration (20 Hz) IR proximity (1 Hz)	Post-hoc infrared event-based
MnM [89]†	32	30 min; 1920 × 1080 30 fps	Bounding boxes (30 min, 1 Hz ‡) F-formations (10 min, 1 Hz) Actions (45 min, 1 Hz‡)	Accelerometer (20 Hz) Radio proximity (1 Hz)	Wearable sync via gossiping protocol; Manual inter-modal sync @1 Hz res.
Conflab	48	~ 45 min; 1920 × 1080 60 fps	17 keypoints (16 min, 60 Hz) F-formations (16 min, 1 Hz) Speaking status (16 min, 60 Hz)	Low-freq. audio (1250 Hz) BT proximity (5 Hz) 9-axis IMU (56 Hz)	Hardware sync at acquisition, max latency ~ 13 ms [253]

† Includes self-assessed personality ratings ‡ Upsampled to 20 Hz using Vatic [258] BT: Bluetooth IMU: Inertial Measurement Unit

head poses, at the cost of less portable sensing setups. Notable examples of such in-the-lab studies include seated scenarios, such as the AMI meeting corpus [109], and more recently standing scenarios like the Panoptic Dataset [257]. Both enable the learning of multimodal behavioral dynamics. However, the dynamics of seated, scripted, or role-playing scenarios are different from that of an unconstrained social setting such as ours. In contrast, Conflab moves out of the lab with a more modular and portable multimodal, multisensor solution that scales easily in the wild.

Another approach exploited wearable sensor data to allow for multimodal processing—sensors included 3 or 6 DOF inertial measurement units (IMU); infrared, Bluetooth, or radio sensors to measure proximity; or microphones for speech behavior [88, 89]. While proximity has been used as a proxy of face-to-face interaction [88, 261–264], recent findings highlight significant problems with such an assumption [265]. Such errors can have a significant impact on the machine-perceived experience of an individual, precluding the development of personalized technology. Chalcedony badges used by [89] show more promising results with a radio-based proximity sensor and accelerometer [266], but such data remains insufficient for more downstream tasks due to the relatively low sample (20Hz) and annotation (1Hz) frequency [81]. In light of these challenges in wearable sensing, Conflab features custom-developed Midge sensors that enable more flexible and fine-grained on-device recording. At the same time, Conflab enables researchers in the wearable and ubiquitous computing communities to investigate the benefit of exploiting wearable and multimodal data.

Furthermore, while both SALSA [88] and MatchNMingle [89] capture a multimodal

dataset of a large group of individuals involved in mingling behavior, the inter-modal synchronization is only guaranteed at 1/3 Hz and 1 Hz, respectively. Prior works coped with lower tolerances by computing summary statistics over input windows [80, 81, 128]. While 1 Hz can capture some conversation dynamics [267], it is insufficient to study fine-grained social phenomena such as back-channeling or mimicry that involve far lower latencies [253, Sec. 3.3]. Conflab provides data streams with higher sampling rates, synchronized at acquisition with our method shown to yield a 13 ms latency at worst [253] (see Sec. 4.3). Table 4.1 summarizes the differences between Conflab and other related datasets.

4.3 DATA ACQUISITION

In this section we describe the considerations, design, and supporting community engagement activities for the first instantiation of Conflab at ACM Multimedia 2019 (MM'19), to serve as a template and case study for other similar efforts.

Ecological Validity and Recruitment An often-overlooked but crucial aspect of in-the-wild data collection is the design and ecological validity of the interaction setting [247–249]. To capture natural interactions in a professional setting and encourage mixed levels of status, acquaintance, and motivations to network, we co-designed a networking event with the MM'19 organizers called *Meet the Chairs!* Our event website (<https://conflab.ewi.tudelft.nl/>) served to inform participants about the goals of a community-created dataset, and transparently describe the data collection process (Figure 4.4). During the conference, participants were recruited via word-of-mouth marketing, social media, conference announcements, and the event website. As an additional incentive beyond interacting with the Chairs and participating in a community-driven data endeavor, we provided attendees with post-hoc insights into their networking behavior from the collected wearable sensors data. See Supplementary material for a sample participant report.

Privacy and Ethics The collection and sharing of Conflab is GDPR compliant. The dataset design and process were approved by the Human Research Ethics Committee (HREC) at our institution (TU Delft) and by the conference location's national authorities (France). When registering, all participants provided consent for the recording and sharing of their data. (See the Datasheet in the Appendix for the consent form.) Given the involvement of private human data, Conflab is only available for academic research purposes under an End User License Agreement. Such an *as open as possible and as closed as necessary* ethos

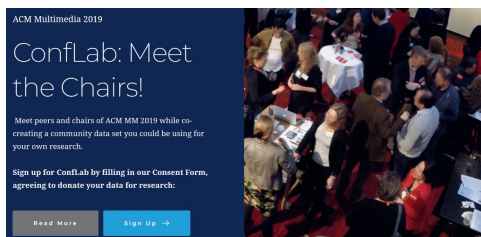


Figure 4.4: Screenshots from the *Conflab: Meet the Chairs!* event website



Figure 4.5: The Midge

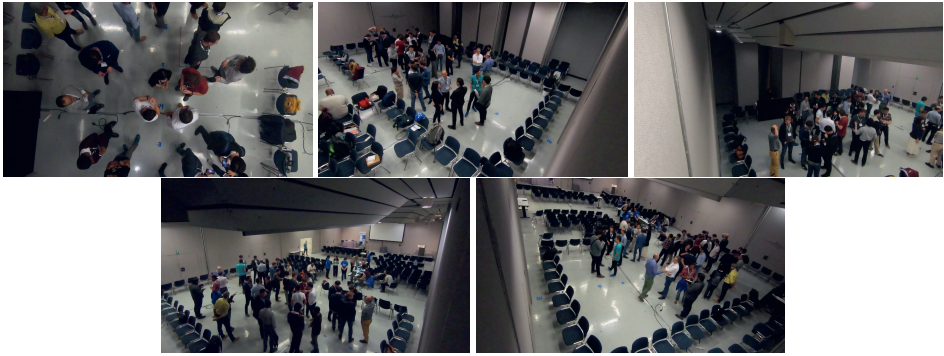


Figure 4.6: Comparing the top-down (top-left, camera 4) and elevated-side camera views (rest). Note how the top-down view is better at mitigating the capture of faces and suffers from fewer occlusions. This allows for a clearer capture of gestures and lower extremities for the most number of people while also preserving privacy.

4

for open science acknowledges the limitation that personal data places on open sharing [268, 269].

Data capture setup Our goal while designing the capture setup was to find the best trade-off between maximizing data fidelity and interfering with the naturalness of the interaction (ecological validity) or violating participant privacy (ethical considerations). Through discussions with the HREC and General Chairs of MM’19 we decided to mitigate the capture of faces, which constitute one of the most sensitive personally-identifiable features. Avoiding the inclusion of faces serves two purposes. First, it safeguards against misuse in downstream tasks with potential negative societal impacts such as harmful surveillance. Such issues have led to the retraction of some person re-identification datasets [251]. Second, it protects the participants who are part of a real research community; since the dataset does not involve role-playing or scripted conversations, the dataset contains their actual behavior. Consequently, we chose an aerial perspective for the video modality (see Figure 4.6). The 10 m × 5 m interaction area was recorded by 14 GoPro Hero 7 Black cameras (60fps, 1080p, Linear, NTSC) [270]. 10 of these were placed directly overhead at a height of ~ 3.5 m at 1 m intervals, with 4 cameras at the corners providing an elevated-side-view perspective. (The HREC has suggested not sharing the elevated-side-view videos due to the presence of faces.) For capturing multimodal data streams, we designed a custom wearable multi-sensor pack called the Midge¹ (see Figure 4.5 for a design render), based on the open-source Rhythm Badge designed for office environments [102]. We improved upon the Rhythm Badge to achieve more fine-grained and flexible data capture (see Appendix 4.D). We designed the Midge in a conference badge form factor for seamless integration. Unlike smartphones, wearable badges allow for a simple *grab-and-go* setup and do not suffer from sensor/firmware differences across models. Popular human behavior datasets are synchronized by maximizing similarity scores around manually identified common events, such as infrared camera detections [88], or speech plosives [271]. While recordings in lab settings can allow for fully wired recording setups, recording in the

¹Documentation and schematics: https://github.com/TUDELFT-SPC-Lab/spcl_midge_hardware

wild requires a distributed wireless solution. We developed a solution to synchronize the cameras and wearable sensors directly at acquisition while significantly lowering the cost of the recording setup [253], making it easier for others to replicate our capture setup. See Appendix 4.D for synchronization and calibration details, and Appendix 4.B for images of the setup.

Data association and participant protocol One consideration for multimodal data recording is the data association problem—how can pixels corresponding to an individual be linked to their other data streams? To this end, we designed a participant registration protocol. Arriving participants were greeted and fitted with a Midge. The ID of the Midge acted as the participant’s identifier. One team member took a picture of the participant while ensuring that both the face of the participant and the ID on the Midge were visible. In practice, it is preferable to avoid this step by using a fully automated multimodal association approach. However, this remains an open research challenge [272, 273]. During the event, participants mingled freely—they were allowed to carry bags or use mobile phones. Conference volunteers helped to fetch drinks for participants. Participants could leave before the end of the one-hour session.

Replicating Data Collection Setup and Community Engagement After the event, we gave a tutorial at MM’19 [274] to demonstrate how our collection setup could be replicated, and to invite conference attendees and event participants to reflect on the broader considerations surrounding privacy-preserving data capture, sharing, and future directions such initiatives could take.

4.4 DATA ANNOTATION

Continuous keypoint annotation Existing datasets of in-the-wild social interactions have mainly focused on localizing subjects via bounding boxes [88, 89]. However, richer information about the social dynamics such as gestures and changes in orientation cannot be retrieved from bounding boxes alone and necessitates the labeling of multiple skeletal keypoints. The typical approach to keypoint annotation involves using tools such as Vatic [258] or CVAT [275] to manually label every N frames followed by interpolating over the rest of the frames. This one-frame-at-a-time annotation procedure makes obtaining keypoint annotations a labor- and cost-intensive process. Moreover, interpolation fails to capture the finer temporal dynamics of the underlying behavior, and reduces the benefits of higher-framerate video capture. Limited by existing tools, no related dataset of in-the-wild human behavior has included time-continuous pose or speaking status annotations.

In contrast, to overcome these issues we collected fine-grained time-continuous annotations of keypoints via a web-based interface implemented as part of the Covfee framework [110]. Here, annotators follow individual joints using their mouse or trackpad while playing the video in their web browser. The playback speed of the video is automatically adjusted using an optical-flow-based technique to enable annotators to follow keypoints continuously without pausing the video. This design enabled easy keypoint labeling in every video frame (60 Hz). We also incorporated a binary *occlusion* flag for every body keypoint. Annotators simultaneously controlled this flag to indicate when a body joint

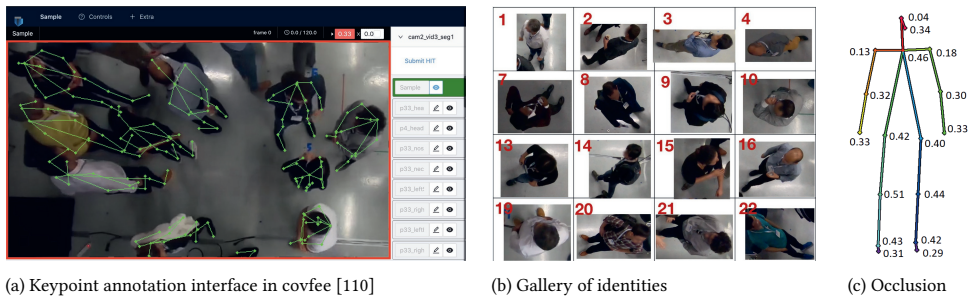


Figure 4.7: Illustration of the body keypoints annotation procedure: (a): our custom time continuous annotation interface; (b): the gallery of person identities used by annotators to identify people in the scene (faces blurred); and (c): the skeleton template with the fraction of occluded frames.

4

was not directly visible. Note that the flag is only an additional confidence indicator; we asked the annotators to label the occluded keypoint using their best estimate of it being within the frame. Our pilot study on the efficacy of Covfee compared to non-continuous annotation via CVAT [275] is presented in [110]. For the pilot annotators, the continuous annotation methodology resulted in a $3\times$ speedup with statistically indifferent error rates.

We chose the top-down camera views for annotation since they suffer from fewer occlusions than the elevated-side views, enabling improved capture of gestures and lower extremities for more people (see Figure 4.6). Given the overlap in the camera views, we annotated keypoints in five of the ten overhead cameras (see Figure 4.1). Note that the same subject could be annotated in multiple cameras due to the overlap in even the five annotated cameras. Videos were split into two-minute segments to ease the annotation procedure. Each segment was annotated by one annotator by tracking the joints of all the people in the scene.

Continuous speaking status annotations Speaking status is a key non-verbal cue for many social interaction analysis tasks [276]. We annotated the binary speaking status of every subject due to its importance as a key feature of social interaction [176, 198, 277–279] and to contribute to the existing community who are working on this task [81, 226, 227]. Action annotations have traditionally been carried out using frame-wise techniques [89], where annotators find the start and end frame of the action of interest using a graphical interface. Given the speed enhancement of continuous annotation, we also annotated speaking status via a continuous technique. We implemented a binary annotation interface as part of Covfee [110]. We asked annotators to press a key when they perceived speaking to be starting or ending. In a pilot study with two annotators, we measured a frame-level agreement (Fleiss’ κ) of 0.552, comparable to previous work [80]. Similar to [89], the annotations were made by watching the video. We provided the annotators with all overhead views to best capture visual behavior.

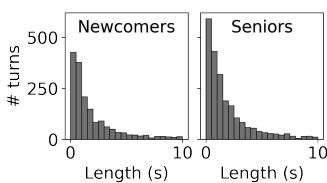
F-formation Annotations Identifying who is likely to have social influence on whom is another important feature for analyzing social behavior. This is operationalized via the theory of F-formations, which are groups of people arranging themselves to converse or socially interact. Similar to prior datasets [88, 89, 250], F-formations group membership

were annotated using an approximation of Kendon’s definition [280]. F-formation stands for Facing formation, a socio-spatial arrangement where people have direct, easy, and equal access while excluding the space from others in the surroundings. The arrangement commonly maintains a convex space in the middle of all the participants (determined by the location and orientation of their lower body). Other spatial arrangements (e.g., side-by-side, L-shaped) are possible, especially for smaller-sized groups of people. Annotations were labeled by one annotator at 1 Hz, following this definition. Since this is a largely objective and common framework for defining F-formations, we deemed it sufficient to obtain one set of annotations. Further, since F-formations may span camera views, we always used a camera that captured each F-formation entirely for annotation.

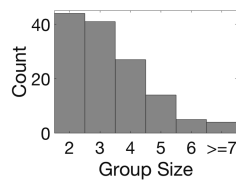
4.5 DATASET STATISTICS

Individual-level statistics Figure 4.7c shows the average occlusion values we obtained from annotators for each of the 17 keypoints. In Figure 4.8a we show the distribution of turn lengths in our speaking status annotations, for both newcomers and veterans, as per their self-reported newcomer status to the conference. We defined a turn as a contiguous segment of positively labeled speaking status, which resulted in a total of 4096 turns annotated.

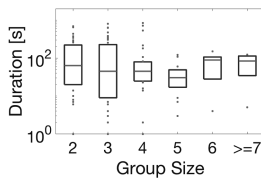
Group-level statistics We found 119 distinct F-formations of size greater than or equal to two, and 38 instances of singletons. Of these, there are 14 F-formations and 2 singletons that include member(s) using the mobile phone. The distributions for group size and duration per group size are shown in Figure 4.8b and Figure 4.8c, respectively. Mean group duration doesn’t seem to be influenced by group size although higher variations are seen at smaller group sizes. The fraction of community newcomers (first-time attending the conference)



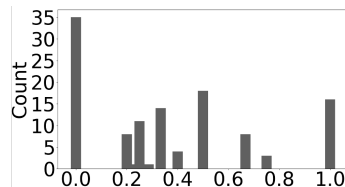
(a) speaking turn lengths



(b) group size



(c) group duration



(d) fraction of newcomers in groups

Figure 4.8: Data distributions for speaking status and conversation groups

Table 4.2: Mask-RCNN results for person bounding box detection and keypoint estimation.

Model	Person Detection			Keypoint Estimation		
	AP ₅₀	AP	AP ₇₅	AP ₅₀ ^{OKS}	AP ^{OKS}	AP ₇₅ ^{OKS}
R50-FPN	73.9	38.9	38.4	45.3	13.5	3.3

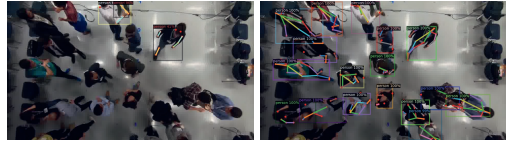


Figure 4.9: Predictions from the Mask-RCNN model; COCO pretrained (left), and Conflab finetuned (right).

in groups is summarized in the histogram in Figure 4.8d. The figure demonstrates two peaks on both sides of the spectrum (i.e., no newcomers vs. all newcomers in the same group). This spread over mixed and non-mixed seniority presents opportunities to study how acquaintance and seniority influence conversation dynamics.

4

4.6 RESEARCH TASKS

We report experimental results on three baseline benchmark tasks: person and keypoints detection, speaking status detection, and F-formation detection. The first task is a fundamental building block for automatically analyzing human social behaviors. The other two demonstrate how learned body keypoints can be used in the behavior analysis pipeline. We chose these benchmarking tasks since they have been commonly studied on other in-the-wild behavior datasets. Code for all benchmark tasks is available at: <https://github.com/TUDELFT-SPC-Lab/conflab>. See the *Uses* section of the Datasheet in the Appendix for a discussion of the broader range of tasks Conflab enables.

4.6.1 PERSON AND KEYPOINTS DETECTION

This benchmark involves the tasks of person detection (identifying bounding boxes) and pose estimation (localizing skeletal keypoints). Since pre-trained SOTA methods struggle with a privacy-sensitive top-down perspective [252] (also see Figure 4.3 and Appendix 4.F.1 for Conflab results), we finetune COCO-pretrained models on our dataset. We used Mask-RCNN [281] (Detectron2 framework [282] implementation) with a ResNet-50 backbone for both tasks for benchmarking. Since keypoint annotations were made per camera, we used four of the overhead cameras for training (Cameras 2, 4, 8, 10) and one for testing (Camera 6). Implementation details are available in Appendix 4.E.1.

Evaluation metrics We evaluated person-detection performance using the standard metrics in the MS-COCO dataset paper [283]. We report average precision (AP) for intersection over union (IoU) thresholds of 0.50 and 0.75, and the mean AP from an IoU range from 0.50 to 0.95 in 0.05 increments. For keypoint detection, we use object keypoint similarity (OKS) [283]. AP^{OKS} is a mean average precision for different OKS thresholds from 0.5 to 0.95.

Results and analyses Table 4.2 summarizes our person detection and joint estimation results. Our baseline achieves 73.9 AP₅₀ in detection and 45.3 AP₅₀^{OKS} in keypoint estimation. Figure 4.9 shows qualitative results from our fine-tuned network. For further insight, we performed several analyses and ablations. In Appendix Table 4.6, we depict the effect of

varying the number of training samples on performance. For training, we use the same four cameras and only vary the number of frames for each camera. We evaluate on the same testing images from camera 6. We find that performance saturates at 16% training samples. We next investigated the effect of increasing training data size by adding specific cameras one at a time. We report results in Appendix Table 4.7. There is a 260% performance gain when first doubling the training samples to 69 k with the addition of camera 4, and a 46% gain when adding another 43 k samples from camera 8. Finally, since the lower body regions suffer from higher occlusion, we experiment with different sections of body for further insight and report results in Appendix Table 4.8.

4.6.2 SPEAKING STATUS DETECTION

In real-life social settings, individual audio recordings can be hard to obtain due to privacy concerns [101]. This has led to exploring other modalities to capture some of the motion characteristics of speaking-related gestures [80, 128]. In this task, we explore the use of body pose and wearable acceleration data for detecting the speaking status of a person in the scene.

Setup We use the SOTA MS-G3D graph neural network for skeleton action recognition [284], pre-trained on Kinetics Skeleton 400. For the acceleration modality, we evaluated three time series classifiers, each of which we trained from scratch: 1D Resnet [285], InceptionTime [285], and Minirocket [286]. We performed late fusion by averaging the scores from both modalities. Like prior work [81, 128], the task was set up as a binary classification problem. We divided our pose (skeleton) tracks into 3-second windows with 1.5 s overlap. A window was labeled positive if more than 50% of the continuous speaking status labels within it were positive. This resulted in an imbalanced dataset of 42882 windows with 29.2% positive labels. Poses were pre-processed for training following [284]. Three of the keypoints (head, and feet tips) were discarded due to not being present in Kinetics. We adapted the network by freezing all layers except for the last fully connected layer and training for five extra epochs. Acceleration readings were not pre-processed, other than by interpolating the original variable-sampling-rate signals to a fixed 50 Hz.

Evaluation Evaluation was carried out via 10-fold cross-validation at the subject level, ensuring that no examples from the test subjects were used in training. We used the area under the ROC curve (AUC) as main evaluation metric to account for the imbalance in the labels.

Results The results in Table 4.3 indicate a better performance from the acceleration-based methods. One possible reason for the lower performance of the pose-based methods is the significant domain shift between Kinetics and Conflab, especially in the camera viewpoint (frontal vs top-down). The acceleration performance is in line with previous work [81]. Multimodal results were slightly higher than acceleration-only results, despite our naive fusion approach, a possible point to improve in future work [287]. Experiments with the rest of the IMU modalities are presented in Appendix 4.F.2.

Table 4.3: ROC AUC and accuracy of skeleton-based, acceleration-based and multimodal speaking status detection (10-fold cross-validation).

Modality	Model	AUC	Acc.
Pose	MS-G3D [288]	0.676	0.677
	InceptionTime [285]	0.798	0.768
Acceleration	Resnet 1D [285]	0.801	0.767
	Minirocket [286]	0.813	0.768
Multimodal	MS-G3D + Minirocket	0.823	0.775

Table 4.4: Average F1 scores for F-formation detection comparing GTCG [256] and GCFE [289] with the effect of different threshold and orientations (standard deviation in parenthesis).

	GTCG		GCFE	
	T=2/3	T=1	T=2/3	T=1
Head	0.51 (0.09)	0.40 (0.12)	0.47 (0.07)	0.31 (0.23)
Shoulder	0.46 (0.11)	0.38 (0.11)	0.56 (0.25)	0.36 (0.16)
Hip	0.45 (0.10)	0.37 (0.12)	0.39 (0.06)	0.25 (0.11)

4.6.3 F-FORMATION DETECTION

4

Setup Like prior work [90, 91, 255, 256], we operationalize interaction groups using the framework of F-formations [280]. We provide performance results for F-formation detection using GTCG [256] and GCFE [289] as a baseline. Recent deep learning methods such as DANTE [255] are not directly applicable since they depend on knowing the number of people in the scene, which is variable for Conflab. We used pre-trained model parameters (reported in the original GTCG and GCFE papers on the Cocktail Party dataset [250]) and tuned a subset of parameters more relevant to Conflab attributes on camera 6. More details can be found in Appendix 4.E.2. We derive three different sets of orientation features from (i) head, (ii) shoulder and (iii) hip keypoints.

Evaluation metrics We use the standard F1 score as evaluation metric for group detection [256, 289]. A group is correctly estimated (true positive) if at least $\lceil T * |G| \rceil$ of the members of group G are correctly identified, and no more than $1 - \lceil T * |G| \rceil$ is incorrectly identified, where T is the tolerance threshold. We report results for $T = \frac{2}{3}$ and $T = 1$ (more strict threshold) in Table 4.4.

Results We show that different results are obtained using different sources of orientations. Different occlusion levels in keypoints due to camera viewpoint may have affected performance. Another factor influencing model performance is that F-formations (which are driven by lower-body orientations [280]) may have multiple conversation floors [278]. Floors are characterized by coordinated speaker turn-taking patterns and influence the head orientations within the group.

4.7 CONCLUSION AND DISCUSSION

Conflab contributes a new concept for real-life data collection in the wild and captures a high-fidelity dataset of mixed levels of acquaintance, seniority, and personal motivations.

Conflab: the dataset We improved upon prior work by providing higher resolution, fidelity, and synchronization across sensor networks. We also carefully designed our social interaction setup to enable a diverse mix of seniority, acquaintanceship, and motivations for mingling. The result is a rich set of 17 body-keypoint annotations of 48 people at 60 Hz from overhead cameras for developing more robust estimation of keypoints, speaking

status, and F-formations for further analyses of more complex socio-relational phenomena. Our benchmark results for these tasks highlight how the improved fidelity of ConfLab can assist in the development of more robust methods for these key tasks. We hope that models trained on ConfLab for localizing keypoints would fill the gap in the cue extraction pipeline, enabling past datasets [89, 90] without articulated pose data to be reinvigorated; this would open the floodgates for more robust analysis of the social phenomena labeled in these other datasets. Finally, our baseline social tasks form the basis for further explorations into downstream prediction tasks of socially-related constructs such as conversation quality [290], dominance [279], rapport [277], influence [291] etc.

ConfLab: the data collection concept To relate an individual's behaviors to trends within their social network, further iterations of ConfLab are needed. These iterations would enable the study of behavioral patterns at different timescales, including multiple interactions in one day, multiple days at a conference, or across distinct conferences. This paper serves as a template for such future ventures. We hope that if the idea of a conference as a living lab gains traction, the effort and cost of data collection can be amortized across different research groups, even involving support from the conference organizers. This *data by the community for the community* ethos can enable the generation of a corpus of related datasets enabling new research questions.

Societal impact ConfLab's long-term vision is to develop technology to assist individuals in navigating social interactions. In this work we have identified choices that maximize data fidelity while upholding ethical best practices: an overhead camera perspective that mitigates identifying faces, recording audio at a low frequency, and using non-intrusive wearable sensors matching a conference badge form factor. We argue this is an essential step towards a long-term goal of developing personalized and socially aware technologies that enhance social experiences. At the same time, such interventions could also affect a community in unintended ways: worsened social satisfaction, lack of agency, stereotyping; or benefit only the members of the community who make use of resulting applications at the expense of the rest. More nefarious uses involve exploiting the data to develop methods that harmfully surveil or profile people. Researchers must consider such inadvertent effects while developing downstream applications. Finally, since we recorded the dataset at a scientific conference and required voluntary participation, there is an implicit selection bias in the population represented in the data. Researchers should be aware that insights resulting from the data may not generalize to the general population.

Empowering users through an agentist rather than structurist approach The analysis of human behavior in social settings has classically taken a more top-down perspective. For instance, the analysis of situated interactions (via only proximity networks) has provided insight into the process of making science in the field of Meta Science [292]. However, while social network science is a well-populated domain, it lacks a more individualized measurement of social behavior: see more discussion of the structure vs. agency debate [293]. Relying on the network science approach jeopardizes an individual's right to technologies that enable free will. We consider the agency in choosing such technologies to be a form of individual harm avoidance. ConfLab provides access to more

than just proximity data about social interactions, enabling the study of context-specific social dynamics. These dynamics are uniquely dependent not only on the individual but also the group they are interacting with [294]. We hope our highlighting of participatory design practices and these value-sensitive design principles promotes social safety in developing socially assistive technologies.

ACKNOWLEDGEMENTS

The authors would like to thank: the ACM Multimedia 2019 General Chairs Martha Larson, Benoit Huet, and Laurent Amsaleg for their support in making the data collection at a major international conference a reality; Bernd Dudzik, Yeshwanth Napoleon, Ruud de Jong, and the venue support staff for their help in setting up the recording on site; Ioannis Protonotarios for the development of the MINGLE Midge badge; Jerry de Vos for improving our Midge Github repository and designing a new case; the participants and student volunteers for the *Meet the Chairs!* event; the Amazon Mechanical Turk workers for their efforts in annotating the dataset; Rich Radke, Martin Atzmueller, Laura Cabrera-Quiros, Alan Hanjalic, and Xucong Zhang for the insightful discussions; Santosh Ilamparuthi for the innumerable discussions and support towards strengthening the ethical soundness of recording and sharing ConfLab; Jan van der Heul for the incredibly responsive support in setting up the 4TU Data repository for ConfLab; and Bart Vastenhouw, Myrthe Tielman, and Catharine Oertel for help with the data sharing; and Musy Ayoub for the word-intelligibility analysis of the low frequency audio.

ConfLab was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated Aspasia Grant, and also by the ACM Multimedia 2019 conference via student helpers, and crane hiring for camera mounting.

Appendices

Conflab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild

4.A HOSTING, LICENSING, AND ORGANIZATION

The dataset is hosted by 4TU.ResearchData, available at <https://doi.org/10.4121/c.6034313>.

The dataset itself is available under restricted access defined by an End-User License Agreement (EULA). The EULA itself is available under a CC0 license. The code (<https://github.com/TUdelft-SPC-Lab/conflab>) for the benchmark baseline tasks, and the schematics and data associated with the design of our custom wearable sensor called the Midge (https://github.com/TUdelft-SPC-Lab/spcl_midge_hardware) are available under the MIT License.

Figure 4.10 on the next page illustrates the organization of the Conflab dataset on 4TU.ResearchData. The components are as follows:

- Annotations (restricted, <https://doi.org/10.4121/20017664>): annotations of pose, speaking status, and F-formations
- Datasheet for Conflab (public, <https://doi.org/10.4121/20017559>): documentation of the dataset following Datasheets for Datasets [295] (see Appendix 4.B)
- EULA (public, <https://doi.org/10.4121/20016194>): End User License Agreement to be signed for requesting access to the restricted components
- Processed-Data (restricted, <https://doi.org/10.4121/20017805>): processed video and wearable sensor used for annotations
- Raw-Data (restricted, <https://doi.org/10.4121/20017748>): raw video and wearable sensor data
- Data Samples (restricted, <https://doi.org/10.4121/20017682>): samples of the sensor, audio, and video data

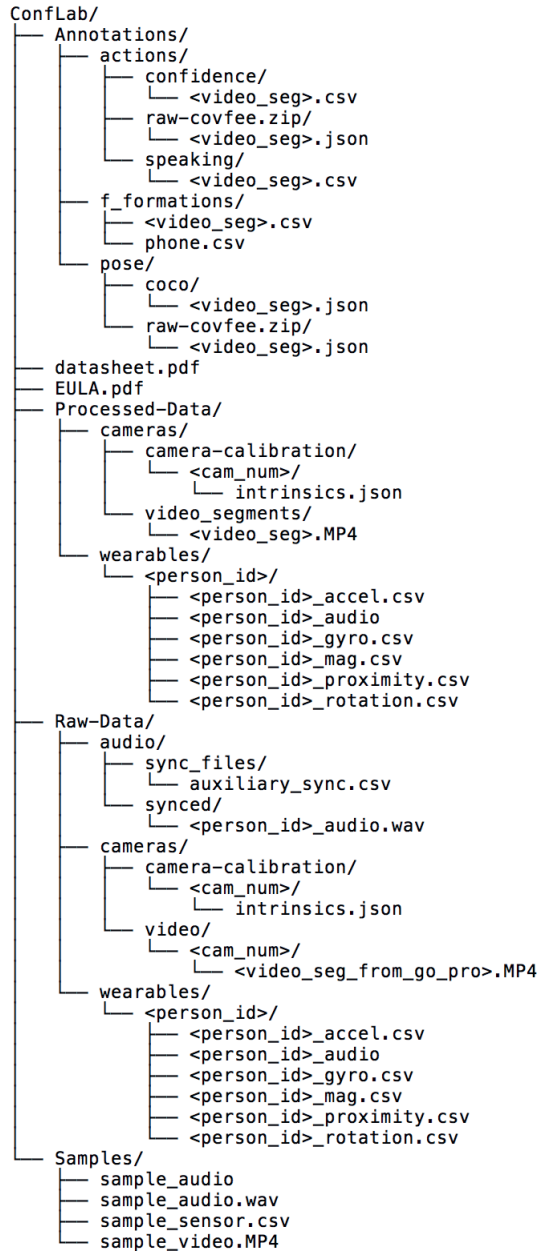


Figure 4.10: File structure of the Conflab dataset

4.B DATASHEET FOR CONFLAB

This document is based on *Datasheets for Datasets* by Gebru *et al.* [295]. Please see the most updated version [here](#).

MOTIVATION

Q. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

There are two broad motivations for creating this dataset: first, to enable the privacy-preserving, multimodal study of *real-life* social conversation dynamics; second, to bring the higher fidelity of wired in-the-lab recording setups to in-the-wild scenarios, enabling the study of *fine time-scale* social dynamics in-the-wild.

We propose the Conference Living Lab (Conflab) with the following goals: (i) a data collection effort that follows a *by the community for the community* ethos: the more volunteers, the more data, (ii) volunteers who potentially use the data can experience first-hand potential privacy and ethical considerations related to sharing their own data, (iii) in light of recent data sourcing issues [254], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

From a technical perspective, closest related datasets (see Table 4.1 in the main paper) suffer from several technical limitations precluding the analysis and modeling of fine-grained social behavior: (i) lack of articulated pose annotations; (ii) a limited number of people in the scene, preventing complex interactions such as group splitting/merging behaviors, and (iii) an inadequate data sampling-rate and synchronization-latency to study time-sensitive social phenomena [253, Sec. 3.3]. This often requires modeling simplifications such as the summarizing of features over rolling windows [80, 81, 128]. On the other hand, past high-fidelity datasets have largely involved role-played or scripted interactions in lab settings, with often a single-group in the scene.

This dataset wasn't created with a specific task in mind, but intends to support a wide variety of multimodal modeling and analysis tasks across research domains (see the *Uses* section).

Q. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Conflab was initiated by the Socially Perceptive Computing Lab, Delft University of Technology in cooperation and support from the general chairs of ACM Multimedia 2019 (Martha Larson, Benoit Huet, and Laurent Amsaleg), Nice, France. Since this dataset was by the community, for the community, members of the Multimedia community contributed as subjects in the dataset.

Q. What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

Conflab was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated Aspasia Grant, and also by the ACM Multimedia 2019 conference via student helpers, and crane hiring for camera mounting.

Q. Any other comments?

None.

COMPOSITION

Q. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset contains multimodal recordings of people interacting during a networking event embedded in an international multimodal machine learning conference.

Overall, the interaction scene contained conversation groups (operationalized as f-formation), composed of individual subjects, each of which had individual data associated to their wearable sensors. The complete interaction scene was additionally captured by overhead cameras. Figure 4.11 shows the structure of these instances and their relationships.

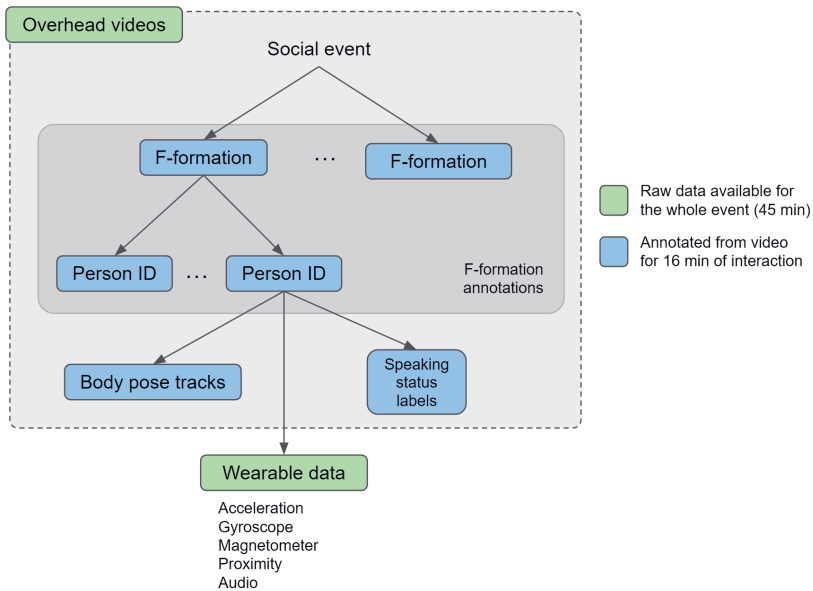


Figure 4.11: Structure of some of the instances in the dataset and their relationships. The interaction space was captured via overhead videos, in which f-formation (conversation groups) were annotated. An F-formation consists of set of people interacting for a variable period of time, and identified via a subject ID. Each person in the F-formation can be associated to their pose (annotated in the videos), their wearable sensor (IMU) data, and their action (speaking status) labels.

Note however that the precise notion of what constitutes an instance in the dataset is very much task-specific. In our baseline tasks we considered the following instances:

Person and Keypoints Detection Frames, containing pose annotations (17 body keypoints per person per frame @60 Hz) from 5 overhead videos (1920 × 1080, 60 fps) for

16 minutes of interaction.

Speaking Status Detection Windows (3 seconds) of wearable sensor data and speaking status annotations (60 Hz) extracted from each subject's data.

F-formations Operationalized conversation groups, annotated at 1 Hz from the 16 minutes of annotated data, and the pose data associated to the people in the F-formation.

Q. How many instances are there in total (of each type, if appropriate)?

The notion of instance is very much dependent on how a user intends to use the data. Regarding the instances in Figure 4.11, our full dataset consist of 45 minutes of:

Video recordings from 10 overhead cameras placed over the interaction area. Five of these videos, enough to cover the complete interaction area, were used in annotation.

Individual wearable sensor data For the 48 subjects in the interaction area, a chest-worn conference-type badge recorded: audio (1250 Hz), and Inertial Measurement Unit (IMU) readings (accelerometer @ 56 Hz, gyroscope @56 Hz, magnetometer @56 Hz and Bluetooth RSSI-based proximity @5 Hz)

Conference experience label For each of the 48 subjects, an associated self-report label indicating whether it was their first time in the conference.

The instances in the annotated 16 minutes segment out of the 45 minutes of interaction contain:

2D body poses For each of the 48 subjects, full body pose tracks annotated at 60Hz (17 keypoints per person). These were annotated using 5 of the 10 overhead cameras due to the significant overlap in views (cameras 2, 4, 6, 8, and 10). Annotations were done separately for each camera by annotating all of the people visible in each video, for each of the 5 cameras, and tagged with a participant ID. We made use of a novel continuous technique for annotation of keypoints. We chose this approach via a pilot study with 3 annotators, comparing our technique to annotations done using the non-continuous CVAT tool. We found no statistically significant differences in errors per-frame (as measured using Mean Squared Error across annotators), despite a 3x speed-up in annotation time in the continuous condition. The details of the technique and this pilot study can be found in [110].

Speaking status annotations For each of the 48 subjects, these include a) a binary signal (60 Hz) indicating whether the person is perceived to be speaking or not; b) continuous confidence value (60 Hz) indicating the degree of confidence of the annotator in their speaking status assessment. These annotations were done without access to audio due to issues with the synchronization of the audio recordings at the time of annotation. The confidence assessment is therefore largely based on the visibility of the target person and their speaking-associated gestures (eg. occlusion, orientation w.r.t. camera, visibility of the face)? We measured inter-annotator agreement for speaking status in a pilot where two annotators labeled three data subjects for 2 minutes each. We measured a frame-level agreement (Fleiss' κ) of 0.552, comparable to previous work [80].

F-formation annotations These annotations label the conversing groups in the scene following previous work. Each individual belongs to one F-formation at a time or is a singleton in the interaction scene. The membership is binary. The annotations were done by one of the authors at 1 Hz by watching the video. The time-stamped usage of mobile phones are available as auxiliary annotations, which are useful for the study of the role of mobile phone users as associates of F-formations. Since Kendon’s theories date back to before the widespread use of mobile phones, their influence on F-formation membership remains an open question.

In our baseline tasks, which made use of the complete annotated section of the dataset, the instance numbers were the following:

Person and Keypoints Detection 119k frames (60fps) containing 1967k person instances (poses) in total, from 48 subjects recorded in 5 cameras (16 minutes of annotated segment).

Speaking Status Detection 42884 3-second windows, extracted from the 48 participants’ wearable data and speaking status annotations.

F-formations 119 conversation groups. Details are in Section 4.5.

Q. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The participants in our data collection are a sample of the conference attendees. Participants were recruited via the conference website, social media posting, and approaching them in person during the conference. Because participation in such a data collection can only be voluntary, the sample was not pre-designed and may not be representative of the larger set. Additionally, 16 minutes of sensor data has been annotated for keypoints, speaking status and F-formations out of the total of 45 minutes recorded. The remaining part (across all modalities) is provided with no labels. For privacy reasons, the elevated cameras (distinct from the previously mentioned 8 overhead cameras) and also individual frontal headshots that were used for manually associating the video data to the wearable sensor data is not being shared.

Q. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Camera 5 failed early during the recording, but the space underneath it was captured by the adjacent cameras due to the high overlap in the camera field-of-views. Nevertheless we share what was recorded before the failure from camera 5, bringing the total number of cameras to 9.

Q. Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The F-formations, subjects, and their associated data relate as shown in Figure 4.11. These associations are made explicit in the dataset via anonymous subject IDs, associated to pose tracks, speaking status annotations, and wearable sensor data. These same IDs were used to annotate the F-formations.

Pre-existing personal relationships between the subjects were not requested for privacy reasons.

Q. Are there recommended data splits (e.g., training, development/validation, testing)?

Since the dataset can be used to study a variety of tasks, the answer to this question is task dependent. Please refer to our reproducibility details (Appendix 4.G of our associated paper) for information about the splits that we used in our baselines.

Q. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Individual audio Because audio was recorded by a front-facing wearable device worn on the chest, it contains a significant amount of cocktail party noise and cross-contamination from other people in the scene. In our experience this means that automatic speaking status detection is challenging with existing algorithms but manual annotation is possible.

Videos and 2D body poses It is important to consider that the same person may appear in multiple videos at the same time if the person was in view of multiple cameras. Because 2D poses were annotated per video, the same is true of pose annotations. Each skeleton was tagged with a person ID, which should serve to identify such cases when necessary.

Q. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Q. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

The data contains personal data under GDPR in the form of video and audio recordings of subjects. The dataset is shared under an End User License Agreement for research purposes, to ensure that the data is not made public, and to protect the privacy of data subjects.

Q. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Q. Does the dataset relate to people?

Yes, the dataset contains recordings of human subjects.

Q. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Data subjects answered the following questions before the start of the data collection event, after filling in their consent form:

- Is this your first time attending ACM MM?

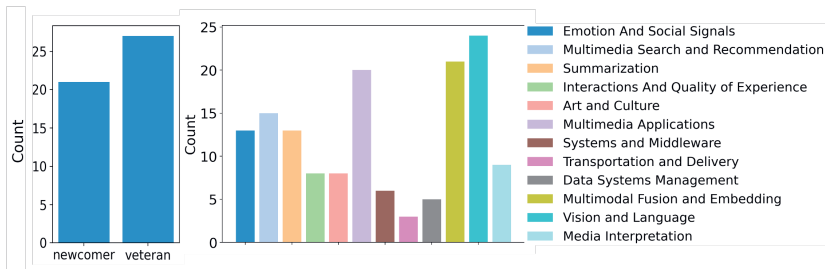


Figure 4.12: Distribution of participant seniority (left) and research interests (right) in percentage.

4

- Select the area(s) that describes best your research interest(s) in recent years. See <https://acmmm.org/call-for-papers/> for descriptions of each theme.

Figure 4.12 shows the distribution of the responses / populations.

Q. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

We do not share any directly identifiable information as part of the dataset. However, individuals may be identified in the video recordings if the observer knows the participants in the recordings personally. Otherwise, individuals in the dataset may potentially be identified in combination with publicly available pictures or videos (from conference attendees or conference official photographer) from other media from the conference the dataset was recorded at. In any case, re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

Q. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

We did not request any such information from data participants. Here, the ACM Multimedia '19 General Chair Martha Larson also helped advocate on behalf of the attendees during the survey-design stage. As a result of these discussions, information such as participant gender, ethnicity, or country of origin was not asked.

Q. Any other comments?

None.

COLLECTION

Q. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The collected data is directly observable, containing video recordings, low-frequency audio recordings and wearable sensing signals (inertial motion unit (IMU) and Bluetooth

proximity sensors) of individuals in the interaction scenes. Accompanying data includes self-reported binary categorization of experience level which is available upon request from the authors. The self-reported interests categories are not shared because of privacy concerns.

Video recordings capture the whole interaction floor where the association from multi-modal data to individual is done manually by annotators by referring to frontal (not-shared) and overhead views. The rest of the data was acquired from the wearable sensing badges, which is person-specific (i.e., no participant shared the device). Video and audio data were verified in playback. Wearable sensing data was verified through plots after parsing.

Q. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the s was created. Finally, list when the dataset was first published.

All data was collected on October 24, 2019, except the self-reported experience level and research interest topics which are either obtained on the same day or not more than one week before the data collection day. This time frame matches the creation time frame of the data association for wearable sensing data. Video data was associated with individual during annotation stage (2020-2021), but all information used for association was obtained on the data collection day.

Q. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

To record videos, we used 14 GoPro Hero 7 Black cameras. The wearable sensor hardware has been documented and open-sourced at https://github.com/TUDeft-SPC-Lab/spl_midge_hardware. The validation of the sensors was completed through an external contractor engineer. The data collection software was documented and published in [110], which includes validation of the system. These hardwares and mechanisms have been open-sourced along with their respective publication.

The synchronization setup for data collection (intramodal and intermodal) was documented and published in [253], which includes validation of the system.

To lend the reader further insight into the process of setting up the recording of such datasets in-the-wild, we share images of our process in Figure 4.13.

Q. What was the resource cost of collecting the data?

The resources required to run this first edition of ConfLab include equipment, logistics, and travel costs. Table 4.5 shows the full breakdown of the costs. The equipment expenses are fixed one-time costs since the same equipment can be used for future iterations of ConfLab. The on-site costs at the conference venue were toward renting a crane for a day to mount the cameras on a scaffold on the ceiling. We have open-sourced the Midge (our custom wearable) schematics so that others don't need to spend on the design and development.

No additional energy consumption was incurred for collecting the data. However, the ancillary activities (e.g., flights, accommodation) resulted in energy consumption. Flights from the Netherlands to France round-trip for six passengers results in 1020 kg carbon emissions. Accommodation for six members resulted in 22 kWh energy consumption.

Q. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

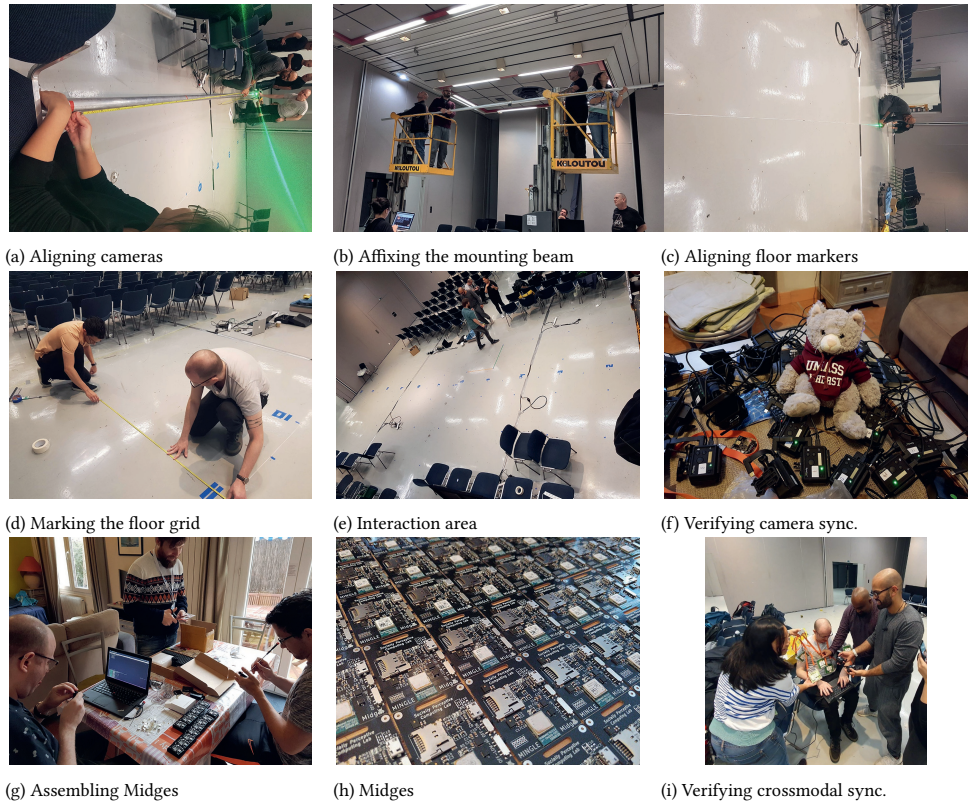


Figure 4.13: Illustrating the process of setting up the data recording.

Conflab contains both annotated and unannotated segments of multi-modal data. The segment where the articulated pose and speaking status were annotated is selected to maximize crowd density in the scenes. The annotated segment is 16 minutes; the whole set is roughly 1 hour of recordings.

Q. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The Conflab dataset was captured during a special social event called *Meet the Chairs!* at an international conference on signal processing and machine learning. Newcomers and old-timers to the conference freely donated their social behaviour data as part of a *by the community, for the community* data collection effort. Aside from the chance to meet the chairs and create a community dataset, the attendees also received a personalised report of their social behaviour from the wearable sensors (see Appendix 4.C) Conference student volunteers were involved in assisting the set-up of the event. Conference organizers (mentioned in the *Motivation* section) assisted in connecting us with conference venue contacts to mount our technical set-ups in the room. Volunteers and conference organizers were not paid by us. Conference venue contacts were paid by the conference organizers.

Table 4.5: Itemized costs associated with recording Conflab

Item	Cost (USD)
Travel (total for 6 people)	
Flights	1800
Accommodation	1500
Equipment (one time)	
Mounting scaffold	2000
14 × GoPro Hero 7 Black	4900
Designing the Midge (custom wearable, now made open source)	26000
110 × Midges (boards, batteries, 4 GB sd cards, cases)	3660
Multimodal synchronization setup	730
Annotations	8000
Computational cost for experiments	500

Data annotations were completed by crowdsourced workers. The crowdsourced workers were paid \$0.20 for qualification assignment (note that typically requesters do not pay for qualification tasks). Depending on the submitted results, workers earn qualification to access of the actual tasks. The annotation tasks were categorized into low-effort (\$150), medium-effort (\$300), and high-effort (\$450), corresponding to the amount of estimated time each would take. The duration of the tasks was determined by the crowd density and through timing of the pilot studies. The average hourly payment to workers is around \$8.

Q. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The data collection was approved by the Human Research Ethics Committee (HREC) of our university (Delft University of Technology), which reviews all research involving human subjects. The data collection protocol is also compliant to the conference location's national authorities (France). The review process included addressing privacy concerns to ensure compliance with GDPR and university guidelines, review of our informed consent form, data management plan, and end user license agreement for the dataset and a safety check of our custom wearable devices.

Q. Does the dataset relate to people?

Yes.

Q. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We collected the data from individuals directly.

Q. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The individuals were notified about the data collection and their participation is voluntary. The data collection was staged at an event called *Meet the Chairs* at ACM MM 2019. The Conflab web page (<https://conflab.ewi.tudelft.nl/>) served to communicate the aim of the

How it Works

- 1. Sign up**
Sign up for ConFlab by filling in our [Informed Consent Form](#), where you agree to take part and to donating your data for research purposes. Still not sure? Read our [FAQ](#).
- 2. Meet the Chairs on Thursday**
Our [data collection](#) will take place on Thursday from 16:30 to 17:30 at the Rhodes hall. After you arrive, we will check if you agreed to donate [your data](#) for research. You fill in a short survey about your research interests and experience with MM. We will give you our newly designed [MINGLE Midge](#) to be worn around your neck. After that you are free to meet peers and the conference's organizational chairs at this event.
- 3. Tutorial and Debrief on Friday**
Join us in learning more about the science and technology behind ConFlab including discussions on privacy, ethics, & data sharing.
- 4. Research and the Future**
Your data will help progress research on social interaction analysis in the wild. It will be shared in a pseudonymised form with the research community under an [Ethical License Agreement](#) to be only used for non-commercial and non-governmental research. We also hope to use it for setting future grand challenges.

What data are you contributing?
We will be collecting the following data as part of this event

- Acceleration and proximity**
Our newly designed MINGLE Midge wearable device records acceleration and proximity during your interactions. Acceleration readings can be used to infer some of your actions like walking and gesturing. It is worn around the neck like a conference badge.
- Video**
Overhead cameras will be mounted to capture the interaction. These videos will be used to annotate behaviour and for detection of social actions like speaking or of conversational groups.
- Low-frequency audio**
The Mingle MIDGE will also record low-frequency audio. This low frequency is enough for recognizing if you are speaking, but not enough to understand the content of your speech, giving us valuable information without compromising your privacy. Example audio:
- Survey measures**
Your research interests and level of experience within the MM community will be linked to the data above via a numerical identifier.

Figure 4.14: Screenshots of the ConFlab web-page used for participant recruitment and registration.

event, what was being recorded, and how participants could sign up. This allowed us to embed the informed consent into this framework so we could keep track of sign ups. See Figure 4.14 for screenshots. This event website was also shared by the conference organizers and chairs (<https://2019.acmmm.org/conflab-meet-the-chairs/index.html>).

Q. Did the individuals in question consent to the collection and use of their data?

If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

All the individuals who participated in the data collection gave their consent by signing a consent form. A copy of the form is attached below in Figure 4.15.

Q. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Yes, the consenting individuals were informed about the possibility of revoking access to their data within a period of 3 months after the data collection experiment, and not after that. The description is included in the consent form.

Q. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

No.

Q. Any other comments?

None.

Declaration of Informed Consent for ConFLab at ACM MM 2019

To take part in this experiment, you must have read the following consent form and agreed to all the points described below. These data will be treated confidentially and will never be linked with your identity or personal information.

By signing, you agree to participate on ConFLab: Meet the Chairs! under the following conditions:

1. During the Meet the Chairs event, we will provide you with the MINGLE Midge sensor to be hung around your neck or clipped to your clothing (we will inform you which you must do at the moment the device is given to you). This device contains a low-power radio (emitter and receiver) for measuring proximity at 5 Hz and ensuring intra-modal synchronization, and an inertial measurement unit (IMU) for measuring body movement. It also records low-frequency audio at a maximum frequency of 2000Hz. A frequency will be chosen that we deem appropriate for detecting speaking status but not enough to recover the content of the conversation. The device has been inspected and deemed safe by a Health Safety and Environment advisor. During operation, the node will record acceleration, angular velocity, orientation, magnetic forces, proximity to other MINGLE Midge wearers, and low-frequency audio in its internal storage.
2. During the experiment, we will be recording video images via cameras installed on the ceiling above the area where you will be interacting, both in top-down and elevated side view. These videos will be treated confidentially and will never be linked to your identity or personal information but we will link your location in the images with the recordings of your MINGLE Midge. To protect your identity, only the top-down videos, where faces are less identifiable, will be shared with other researchers. However, we cannot guarantee that you cannot be identified from the video images.
3. To link your video data with your MINGLE Midge data, we use a camera to record a frontal video of you stating or showing your numerical identifier to the camera. The data from the frontal camera will not be shared.
4. The identity of your MINGLE Midge will be linked to the numeric identifier that you will receive when entering the room where the experiment is performed. This allows us to ensure that everybody who is recorded has agreed with this declaration.
5. Your recordings will be linked to the answers of the survey that you will be asked to fill during the event via a numerical identifier. They will also be linked to the following information from your ACM MM 2019 registration:
 - a. years of experience in the field
 - b. research interests
6. The recorded data will not be made freely available to the general public. The data may be shared with other researchers in the research community, only in the case of research that is substantially similar in purpose to the goal of this research project (analysis of community/network dynamics, analysis of social interaction in mingling scenarios) and only if these parties comply with the European Union General Data Protection Regulation (GDPR). Any researchers requesting access to the data will be required to sign an End-User License Agreement (EULA) agreeing to keep the data private and to the responsible use of the data as described in point 6, as well as compliance with the GDPR.
7. You understand that your participation in this experiment is voluntary. You have the right to withdraw from the experiment at any time during its execution. You may have access to your data if you request it. You have the right to the deletion of your data during a period of 3 months after the experiment, but not after this period. If you request deletion, we will ensure that your data is removed from the collection. In the case of video data, we will ensure that your face is anonymized/blurred in all videos.
8. In all cases, excerpts of the data that are used in research publications or presentations will be anonymized. This means that your identity will not be linked to your data, and we will ensure that your face is blurred in the images. The anonymized data may be presented in the following ways:
 - Screenshots of the videos may be published in scientific publications.
 - We may use short excerpts of the videos in scientific presentations.
 - In the event that the experiments are of interest to the press, anonymized excerpts of the data may be distributed to the media (e.g. Newspapers, TV).

I agree to participate in ConFLab and to the sharing of my data:

I agree

Name of Participant:

Signature of participant:

4

Figure 4.15: Consent form signed by each participant in the data collection.

PREPROCESSING / CLEANING / LABELING

Q. Was any preprocessing/cleaning/labeling of the data done(e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

We did not pre-process the signals obtained from the wearable devices or cameras. The only exception is the audio data. Due to a hardware malfunction (this is resolved for the Midges by using different SD cards), the audio needed to be post-processed in order to synchronize it with the other modalities. The synchronization against other modalities was manually checked.

Labeling of the dataset was done as explained in the *Composition* section.

Q. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The dataset is separated into raw data and the post processed data. For the audio, the original raw data is not suitable for most use cases due to the mentioned synchronization issue. So we share the synchronized version in the raw part of the repository.

Q. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The processing / fixing of the audio files did not require special software.

The annotation of keypoints and speaking status was done by making use of the Covfee framework: <https://josedvq.github.io/covfee/>

Q. Any other comments?

None.

USES

Q. Has the dataset been used for any tasks already? If so, please provide a description.

In the main paper, we have benchmarked three baseline tasks: person and keypoints detection, speaking status detection, and F-formation detection. The first task is a fundamental building block for automatically analyzing human social behaviors. The other two demonstrate how learned body keypoints can be used in the behavior analysis pipeline for inferring more socially related phenomena. We chose these benchmarking tasks since they have been studied on other in-the-wild behavior datasets.

Q. Is there a repository that links to any or all papers or systems that use the dataset?

None at the time of writing of the paper.

Q. What (other) tasks could the dataset be used for?

Given the richness and the unscripted open-ended nature of the social interactions, ConfLab can be used for many other tasks.

Forecasting, causal relationship discovery Recently, tasks pertaining to the forecasting low-level social cues in conversations have been receiving increased attention from the community [294, 296]. The real-life nature of ConfLab along with the increased data and annotation fidelity can prove a valuable resource for such tasks. Similarly, ConfLab can also be used for efforts towards discovering causal relationships between social behaviors [297].

Data Association. A crucial assumption made in many former multimodal datasets [88, 89, 257] is that the association of video data to the wearable modality can be manually performed. Few works [272, 273] have tried to address this issue but using movement cues alone to associate the modalities is challenging as conversing individuals are mostly stationary. This remains a significant and open question for future large scale deployable multimodal systems. One solution may be to annotate more social actions as a form of top-down supervision. However, detecting pose and actions robustly from overhead cameras remains to be solved.

Conversation floor and F-formation estimation Prior analysis on the MatchNMingle dataset has demonstrated that F-formations can contain multiple simultaneous conversations when the F-formations contain a least 4 people [278]. If this is the case for the ConfLab dataset, this may drastically change how F-formations should be labelled (e.g. returning to being a more subjective task [90]) as more time-precise labelling could enable a more nuanced take on F-formation and conversation floor membership over time.

Multi-class social action estimation More annotations resources were focused on speaker status, F-formation, and keypoint estimation. However, there are a wealth of other social actions in the data that could be interesting to combine into a more complex multi-class social action estimation task. Example social actions include drinking, mobile phone use, hand and head gesture types [85, 89].

Estimation and analysis of socially-related phenomena Beyond the modeling of human behavior which is of interest to the Computer Vision and Machine Learning communities, our benchmarked tasks form the basis for further explorations into downstream prediction of socially-related constructs which is of interest to the Social Science and Social Psychology communities. Such constructs include conversation quality [290, 298], dominance [279], rapport [277], and influence [291].

Investigation of novel crossmodal fusion strategies The baseline tasks in our paper rely only on a late fusion strategy. However, Conflab's sub-second expected cross modal latency of ~ 13 ms along with higher sampling rate of features (60 fps video, 56 Hz IMU) opens the gateway for the in-the-wild study of nuanced time-sensitive social behaviors like mimicry and synchrony (for predicting e.g. attraction [57]) which need tolerances as low as 40 ms [253, Sec.3.2]. Prior works coped with lower tolerances by computing summary statistics over input windows [80, 81, 128]. Conflab enables for the first time, the exploration of Multimodal machine learning approaches for social behaviour analysis in these highly dynamic in-the-wild settings [287]. Through the provided annotations Conflab also enables research in the topic of usage of mobile phones in small-group social interactions in-the-wild.

Person attribute estimation Estimating individuals that are newcomers/old timers from the dataset may be possible based on their networking strategies.

Q. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Although Conflab's long-term vision is towards developing technology to assist individuals in navigating social interactions, the data could also affect a community in unintended ways: for instance, cause worsened social satisfaction, a lack of agency, stereotype newcomers and veterans, or benefit only those members of the community who make use of resulting applications at the expense of the rest. More nefarious uses involve exploiting the data for developing methods that harmfully surveil or profile people. Researchers must consider such inadvertent effects must while developing downstream applications. Finally, since we recorded the dataset at a scientific conference and required voluntary participation, there is an implicit selection bias in the population represented in the data. Consequently, researchers using the data should be aware that resulting insights may not generalize to the general population.

Q. Are there tasks for which the dataset should not be used? If so, please provide a description.

Beyond the cautionary discussion in the previous question, tasks involving the re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

Q. Any other comments?

None.

DISTRIBUTION

Q. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset is available for third parties outside of Delft University of Technology to use for academic research purposes subject signing and approval of our End User License Agreement. The dataset will be hosted by 4TU.ResearchData (see the Maintenance section for description of the 4TU entity).

Q. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed via the 4TU.ResearchData user interface where the data can be downloaded. The dataset has a DOI: <https://doi.org/10.4121/c.6034313>

Q. When will the dataset be distributed?

The dataset has been available since June 9, 2022.

Q. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed under a restricted copyleft license, specified within our End User License Agreement, accessible through the 4TU.ResearchData dataset website. No fees are associated with the license.

Q. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Q. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

The terms of our EULA and the European General Data Protection Regulations (GDPR) apply.

Any other comments?

None.

MAINTENANCE

Q. Who is supporting/hosting/maintaining the dataset?

The dataset is hosted by 4TU.ResearchData (https://www.4tu.nl/en/about_4tu/), and supported and maintained by The Socially Perceptive Computing Lab at TUDelft.

Q. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Via email: SPCLabDatasets-insy@tudelft.nl.

Q. Is there an erratum?

No.

Q. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Updates will be done as needed as opposed to periodically. Instances could be deleted, added, or corrected. The updates will be posted on the 4TU.ResearchData dataset website.

Q. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No limits were communicated to our data participants.

Q. Will older versions of the dataset continue to be supported/hosted/maintained?

If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Only the latest version of the dataset will be maintained. If applicable, we will also host older versions of the data, accessible through the 4TU.ResearchData website.

Q. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We are open to contributions to the dataset. In accordance with our End User License Agreement, contributions should be made available, indicating if there are any restrictions on their contribution. We encourage the potential contributors to contact us to discuss how they wish to be attributed (e.g. citation of a paper or repository related to code/annotations). After finalizing the attribution discussion, we can add the attribution as an update following the same process explained above.

4.C SAMPLE PARTICIPANT REPORT

ACMMM 19 - ConFlab Report

Socially Perceptive Computing Lab - Delft University of Technology

Conflab: Meet the Chairs!

While you were at ACM MM in Nice earlier this year, you had participated in our event called Conflab: Meet the Chairs! We want to thank you again for being part of our data collection initiative and contributing to the effort of understanding more about human behaviors and conference experience.

We thought you might be curious about some basic statistics that we have extracted from the collected data. You can find below some general information about all the event participants and some personal information particular to you. Please keep in mind that 1) these are preliminary analyses that we have performed and there could be errors in our estimations, and 2) to protect your privacy, these results are only available to you.

General information about Conflab participants

When you signed up, we had asked 1) if this was your first time at ACM MM and 2) your research interests (multi-select multiple choice). We had a total of 48 participants. You can see below the statistics over all 48 people.

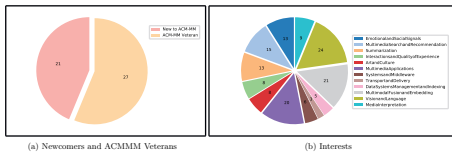


Figure 1: Statistics of Conflab participants

1

Your networking behaviour - Bluetooth

Here we estimate how many people you have interacted with throughout the event. Our sensors record RSSI values and we set a single threshold for eliminating values corresponding to large physical distances that we do not consider as possible for face-to-face social interactions. We define the criterion of an interaction to be: 1) pairwise RSSI values below -55, and 2) pairwise proximity pings of at least 35 counted within a 1-minute window (sampling rate: 1Hz).

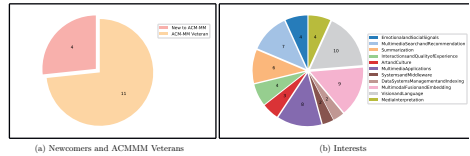


Figure 2: Statistics of people you interacted with

In Figure 2a, the breakdown of the types of people you have interacted with is shown. In Figure 2b, you will find the interests breakdown of everyone you have interacted with. Figure 3 shows the distribution of the number of participants you interacted with. You will find yourself in the red bin; the x-axis says how many people you have interacted with and the y-axis says how many others had the same numbers as you.

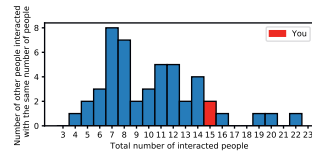


Figure 3: Distribution of the numbers of people participants interacted with

2

Your movement behavior - accelerometer

Here we estimate your motion behavior based on the accelerometer signal. Our sensors record tri-axial accelerometer values and we quantify the amount of motion by calculating the magnitude of the values of all 3 axes. We process the accelerometer data to separate movement and gravitational components of the signals based on a previous approach (Euclidean Norm Minus One [1]). For ease of visualization, we averaged the magnitude of acceleration over 30-second windows. You can see in Figure 4 your personal acceleration magnitude over time, as well as the mean and standard deviation values of acceleration magnitude for all participants over time.

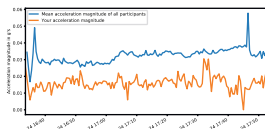


Figure 4: Acceleration magnitudes

Your speech behaviour - low-frequency audio

Here we estimate the amount of time you spoke. We first calculate the envelope of the low-frequency audio signal by taking the absolute value. Then, we apply a moving mean operator to the signal. By manually observing the signals of multiple participants, we selected a threshold to identify the speaking parts of the signal. We then further process the binary stream by filling the gaps between continuous speaking regions and eliminating speech regions that are smaller than a predefined threshold. Figure 5a and 5b show your percentage of speaking during the event and how you compare to the rest of the participants, respectively.

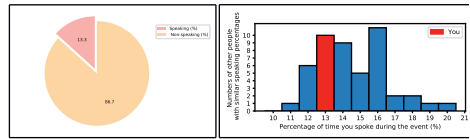


Figure 5: Your speaking behaviour

And that's it from the Socially Perceptive Computing Lab for now!

Note that for us, these analyses are just the starting point for estimating socially relevant behaviours. To do this more robustly and using more complex approaches is one of the reasons why we plan to share the data in next year or so. Maybe you are also curious to develop your own estimation techniques.

Finally, we welcome feedback on what other analyses that you are interested in, technical approaches, how to display your data better, your participatory experience, and any comments or advice that you might have for us. Please feel free to reply to this email or write to one of us directly.

Thanks again for your interest and we hope to see you again in the future!

[1] Bakranji, Kishan, et al. "Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches." *PLoS one* 11.10 (2016): e0164045.

3

4

Figure 4.16: Sample post-hoc report sent to each participant of ConFlab. The report contains insights into the participant's networking behavior from the collected wearable-sensors data. This insight served as an additional incentive to participate in ConFlab, beyond interacting with the Chairs and contributing to a community-driven data endeavor (see main paper Section 4.3).

4.D DATA CAPTURE SETUP DETAILS

The Midge We improved upon the Rhythm Badge in three ways towards enabling more fine-grained and flexible data capture: (i) enabling full audio recording with a frequency up to 48 KHz, with an on-board switch to allow physical selection between high and low frequency capture directly at acquisition; (ii) adding a 9-axis Inertial Measurement Unit (IMU) with an on-board Digital Motion Processor (DMP) to record orientation; and (iii) an on-board SD card to directly store raw data, avoiding issues related to packet loss during wireless data transfer required by the Rhythm Badge. IMUs combine three tri-axial sensors: an accelerometer, a gyroscope, and a magnetometer. These measure acceleration, orientation, and angular rates respectively. These sensor measurements are combined on-chip by a Digital Motion Processor. Rough proximity estimation is performed by measuring the Received Signal Strength Indicator (RSSI) for Bluetooth packets broadcast every second (1 Hz) by every Midge. During the event, IMUs were set to record at 50 Hz. We recorded audio at 1250 Hz to mitigate extraction of verbal content while still ensuring robustness to cocktail-party noise.

Wireless synchronization at acquisition The central idea for our synchronization approach involves using a common Network Time Protocol (NTP) signal as reference for the camera and wearables sub-networks. The set-up achieved a cross-modal latency of 13 ms at worst, which is well below the 40 ms latency tolerance suitable for behavior research in our setting [253, Sec. 3.3]. Additionally, our synchronization approach allowed for dynamic addition of sensors to the network while still obtaining synchronized data streams. This is crucial in extreme in-the-wild events where some participants might arrive late.

Sensor calibration For computing the camera extrinsics, we marked a grid of $1\text{ m} \times 1\text{ m}$ squares in tape across the interaction area floor. We ensured line alignment and right angles using a laser level tool (STANLEY Cross90). For computing the camera intrinsics, we used the OpenCV asymmetric circles grid pattern [299]. The calibration was performed using the Idiap multi camera calibration suite [300]. All wearable sensors include one TDK InvenSense ICM-20948 IMU [301] unit that provides run time calibration. To establish a correspondence with the camera frame of reference, the sensors were lined up against a common reference-line visible in the cameras to acquire an alignment so that the camera data can offer drift and bias correction for the wearable sensors.

4.E IMPLEMENTATION DETAILS

4.E.1 PERSON AND KEYPOINT DETECTION MODELS

Data cleaning A few frames contained some incorrectly labeled keypoints, a product of annotation errors like mis-assignment of participant IDs. We removed these using a threshold on the proximity to other keypoints of the same person. Further, in some cases, a person might be partially outside a camera's field of view. For the person detection task, we compute the bounding box from the keypoint ground-truth annotations. If more than half the body (50% keypoints) is missing in the frame so that e.g. only their legs are visible (see top of Figure 4.7a), we don't consider the person for that frame in the person detection

experiments. Note that due to the significant overlap between the camera views, the person would be considered for the corresponding frame in the next camera. If they move back into the original view, we again take them into consideration for the original camera for the corresponding frame. Moreover, if there are more than 10% missing keypoints across all people in an image, we also discard that image from the experiment. This preprocessing resulted in a training set with 112k frames (1809k person instances) and a test set with 7k frames (158k person instances).

Training We resized the images to 960×540 , and augmented the data by randomizing brightness and horizontal flips. The learning rate was set to 0.02 and batch size to 4. We trained the models for 50 k iterations, using the COCO-pretrained weights for initialization. All hyper-parameters were chosen based on the performance on a separate hold-out camera chosen as validation set. During training, any missing ground-truth keypoints (resulting from the person being partially outside the camera’s view for instance) are ignored during back-propagation.

4.E.2 F-FORMATION DETECTION

Data Cleaning Because keypoint annotations of the subjects are based on camera view and that the F-formation clustering methods cannot group subjects that do not exist under one camera view (e.g., when there are more identities than in associated ground truths), we processed the ground truth also based on camera number. This filtering pre-processing was decided based on the best camera view of the F-formations.

Feature extraction The required features of GCFF and GTCG include location and orientation of the subjects. We used the X and Y position of subjects’ head (as it is the most visible from the top-down view) for location, and extracted orientations for head, shoulders and hips. The orientations are calculated based on corresponding vectors determined by head and nose keypoints, left and right shoulder keypoints, and left and right hip keypoints, respectively.

Training We used pre-trained parameters for field of view (FoV) and frustum aperture (GTCG) and minimum description length (GCFF), provided in these models trained on the Cocktail Party. FOV and aperture are related to human eye gaze and head anatomical constraints reported by [302], and hence not dataset specific. The minimum description length is an initialized prior dictated by the same form of the Akaike Information Criterion, and becomes part of the optimization formulation. We tuned parameters such as frustum length (GTCG) and stride (GCFF) to account for average interpersonal distance in Conflab based on Camera 6, as they vary across different datasets.

4.F ADDITIONAL RESULTS

4.F.1 PERSON AND KEYPOINTS DETECTION

Predictions from pretrained SOTA models Figure 4.17 shows predictions from SOTA human keypoint estimation models, namely, RSN [246], MSPN[303], HigherHRNet [304], and HourglassAENet [305], for the testing images of the Conflab dataset. Note that RSN

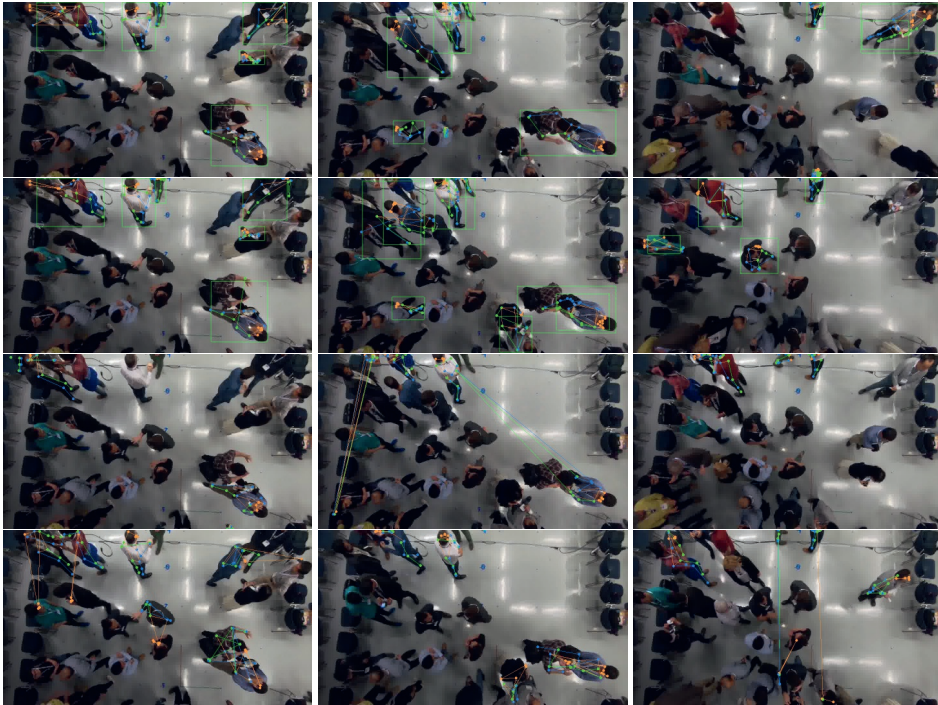


Figure 4.17: Results from pre-trained keypoint detection models. From top to bottom - predictions from RSN [246], MSPN[303], HigherHRNet [304], and HourglassAENet [305]. Results show that *SOTA 2D body keypoint detection models fail to capture the body keypoints in the Conflab dataset.*

and MSPN are top-down networks, i.e., they require person bounding boxes to predict the keypoints in each bounding box. We use COCO pretrained faster-RCNN network for bounding box estimation. HigherHRNet and HourglassAENet are bottom-up models, i.e., they directly predict keypoints from the full image. We use publicly available COCO pretrained checkpoints for prediction. The results show that the *state-of-the-arts 2D body keypoint detection models fail to capture the body keypoints in the Conflab dataset.* We infer that training on the dataset (e.g., COCO) that contains mostly side-view images does not work well in top-view images, for which Conflab dataset is important to the community.

Qualitative results from ResNet-50 finetuning Figure 4.18 illustrates more qualitative results from our finetuning experiments. We find that finetuning on our non-invasive top-down camera perspective significantly improves the keypoint estimation performance.

Ablations Tables 4.6 and 4.7 include the results of our experiments investigating the effect of varying the training data size on keypoint detection performance (see main paper Section 4.6.1). In Table 4.8, we show keypoint detection scores for experiments with different number of keypoints. We first focus on the five upper body keypoints: {head, nose, neck, rightShoulder, leftShoulder}. We then additionally considered the torso region

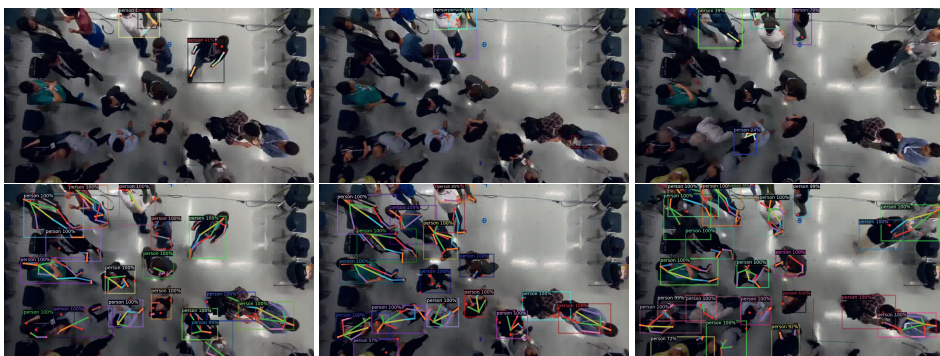


Figure 4.18: Results from (top) COCO pretrained Mask-RCNN model, (bottom) our Conflab finetuned Mask-RCNN model.

4

Table 4.6: Effect of varying % frames from each camera at training on keypoint estimation.

% of training samples	AP_{50}^{OKS}
1.6%	29.0
3.2%	35.9
8%	39.0
16%	44.5
100%	45.3

Table 4.7: Effect of adding all frames from individual cameras to the training set on keypoint estimation.

Train Camera	#(training samples)	AP_{50}^{OKS}
cam 2	34k	8.6
cam 2 + cam 4	69k	31.1
cam 2 + cam 4 + cam 8	112k	45.3

keypoints for a total of nine: {rightElbow, rightWrist, leftElbow, leftWrist}. Finally, we add the hip keypoints {rightHip, leftHip} to the set. The experiments in the main paper are performed with all 17 keypoints. The results show that performance drops slightly when adding the arms keypoints ($5 \rightarrow 9$, AP_{50}^{OKS} and AP^{OKS}), and that the relative gain when adding the hip keypoints ($9 \rightarrow 11$) is lower than when adding the lower body keypoints ($11 \rightarrow 17$, especially AP_{75}^{OKS}). We believe this is largely due to the lower body being more static relative to the arms that move a lot to execute gestures during conversations.

4.F.2 SPEAKING STATUS DETECTION

Experiments with different sensor modalities Table 4.9 displays the results from experiments using specific modalities from our IMUs for the task of speaking status detection. We used the best performing classifier (Minirocket [286]) among the ones tested in Table 4.3. The experiment setup is the same as detailed in Section 4.6.2, and the model is not changed between runs, except for the fact that different modalities may have a different number of input channels.

Table 4.8: Keypoint estimation ablation with keypoints from different body sections: head and shoulders (5), + torso (9), + hips (11), + knees and feet (full 17).

#Keypoints	AP ₅₀ ^{OKS}	AP ^{OKS}	AP ₇₅ ^{OKS}
5	26.6	7.1	1.4
9	26.5	6.9	2.0
11	35.8	9.5	2.2
17	45.3	13.5	3.3

Table 4.9: ROC AUC and accuracy for different sensor modalities from out 9-dof IMU in speaking status detection using the Minirocket classifier [286]. The number of channels in the corresponding modality is indicated in parentheses.

Input Modality	AUC	Accuracy
Acceleration (3)	0.813	0.768
Gyroscope (3)	0.765	0.716
Magnetometer (3)	0.610	0.656
Rotation vector (4)	0.726	0.696
All (13)	0.774	0.739

4.G REPRODUCIBILITY CHECKLIST

4.G.1 PERSON AND KEYPOINTS DETECTION

- Source code link: <https://github.com/TUDeft-SPC-Lab/conflab>
- Data used for training: 112k frames (1809k person instances).
- Pre-processing: See Section 4.4, Appendix 4.E.1.
- How samples were allocated for train/val/test: cameras 2, 4, and 8 are selected for training. For hyperparameter tuning, camera 8 are held out for validation.
- Hyperparameter consideration: We considered learning rates (0.001/0.005/0.05/0.01), number of epochs (10/20/50/100), detection backbone (R50-FPN/R50-C4). Also see Appendix 4.E.1
- Number of evaluation runs: 5
- How experiments were ran: See Section 4.6.1.
- Evaluation metrics: Average precision at different thresholds.
- Results: See Section 4.6.1 and Appendix 4.F.1.
- Computing infrastructure used: All baseline experiments were ran on Nvidia V100 GPU (16GB) with IBM POWER9 Processor.

4.G.2 SPEAKING STATUS DETECTION

- Source code link: <https://github.com/TUDeft-SPC-Lab/conflab>
- Data used for training: 42884 windows (3 seconds), extracted from 48 participants' wearable data and speaking status annotations
- Pre-processing: Data was windowed into 3-second segments (see Section 4.6.2). The source code includes this pre-processing step.

- How samples were allocated for train/val/test: 10-fold cross-validation at the subject level (48 subjects) to test generalization to unseen data subjects. The splits can be reproduced exactly using the source code.
- Hyperparameter considerations: For acceleration-based methods, we used default network hyper-parameters and architectures from their tsai implementation [306]. For the MS-G3D baseline [284], we used default hyperparameters from the authors' implementation. For both, we determined the early stoppage point using a small subset (10%) of the training set.
- Number of evaluation runs: 1 run of 10-fold cross-validation
- How experiments were ran: For each fold, the early stoppage point was first determined using 10% of the training data as validation set and AUC as performance metric. The model at this stoppage point was then applied to the test set for evaluation.
- Evaluation metrics: Area under the ROC curve (AUC)
- Results: See Section 4.6.2
- Computing infrastructure used: Experiments were ran on a personal computer with GPU acceleration (Nvidia RTX3080).

4.G.3 F-FORMATION DETECTION

- Source code link: <https://github.com/TUDeft-SPC-Lab/conflab>
- Data used for training: Camera 6
- Pre-processing: See Section 4.E.2 for data cleaning and feature extraction.
- How samples were allocated for train/val/test: samples from Camera 6 were used to select the best model parameters. The rest are for test (evaluation). However, we note that Table 4.4 shows averaged performance on all cameras to provide a holistic view of the F-formation detection performance on Conflab.
- (Hyper)parameter considerations: Both baseline methods are not deep-learning based and model parameters are interpretable. For GTCG, the parameters are frustum length (275), frustum aperture (160), frustum samples (2000), and sigma for affinity matrix (0.6). For GCFE, the parameters are minimum description length (30000) and stride (70).
- Number of evaluation runs: 1
- How experiments were ran: A total of eight experiments were run for choosing the best parameters, and three for evaluation (for camera 2, 4, and 8). The parameters were chosen based on grid-search. For optimizing frustum length in GTCG, we searched over [170, 195, 220, 245, 275] with 275 being averaged interpersonal distance based on Camera 6. For optimizing stride D in GCFE, we searched over [30, 50, 70].
- Evaluation metrics: F1

- Results: See Section 4.6.3
- Computing infrastructure used: The experiments were run on Linux-based cluster instances on CPU with Matlab 2018a.

5

REWIND DATASET: PRIVACY-PRESERVING SPEAKING STATUS SEGMENTATION FROM MULTIMODAL BODY MOVEMENT SIGNALS IN THE WILD

5

Recognizing speaking in humans is a central task towards understanding social interactions. Ideally, speaking would be detected from individual voice recordings, as done previously for meeting scenarios [109]. However, individual voice recordings are hard to obtain in the wild, especially in crowded mingling scenarios due to cost, logistics, and privacy concerns [95]. As an alternative, machine learning models trained on video and wearable sensor data make it possible to recognize speech by detecting its related gestures in an unobtrusive, privacy-preserving way. These models should ideally be trained using labels obtained from the speech signal. However, existing mingling datasets do not contain high-quality audio recordings. Instead, speaking status labels are often inferred by human annotators from video, without validation of this approach against audio-based ground truth. In this paper, we revisit no-audio speaking status estimation by presenting the first publicly available multimodal dataset with high-quality individual speech recordings of 33 subjects in a professional networking event. We present three baselines for no-audio speaking status segmentation: a) from video, b) from body acceleration (chest-worn accelerometer), c) from body pose tracks. In all cases, we predict a 20Hz binary speaking status signal extracted from the audio, a time resolution not available in previous datasets. In addition to providing the signals and ground truth necessary to evaluate a wide range of speaking status detection methods, the availability of audio in REWIND makes it suitable for cross-modality studies not feasible with previous mingling datasets. Finally, our flexible data consent setup creates new challenges for multimodal systems under missing modalities.

5.1 INTRODUCTION

Detection or segmentation of speaking activity in free-standing social settings is a core necessity in building systems capable of interpreting interactions in everyday situations, from networking events to exchanges around the coffee machine at the office. The analysis of a complex conversational scene where dozens of people stand, walk, form groups, and converse freely (see Fig. 5.1) is of particular interest in fields such as computational social science and social signal processing for the development of socially intelligent systems capable of aid [249]. Segmenting speaking status (ie. binary signal indicating voice activity of a target speaker) with time resolutions that are suitable for indicating back-channels is key because of its utility in downstream tasks where it can be used, for example, in the quantification of individual and group measures of experience in conversation like involvement [52], satisfaction [77], perceived quality [290, 298], or affect [76], and in the forecasting of future events like speaking, gesturing and changes in position and orientation [131, 294].

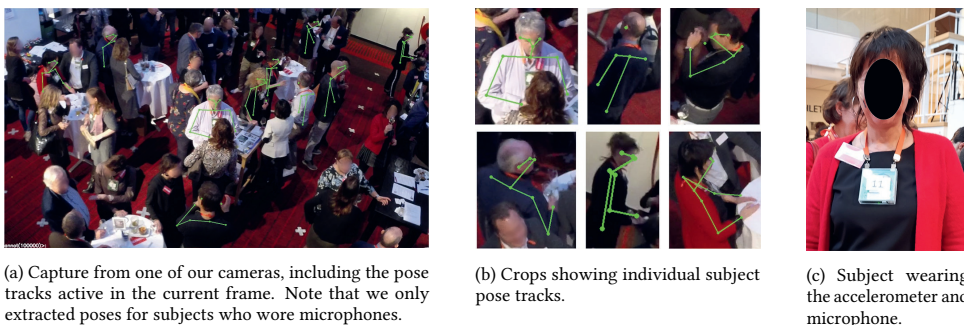
The audio modality is the obvious choice for the measurement of speaking status. High-quality speaking status signals have been obtained from personal head-mounted and directional microphones in seated meetings [109]. However, individual audio is especially hard to acquire in a mingling setting, or crowded conversational scene. Microphone equipment is hard to scale to large dynamic crowded scenes with synchronization guarantees. Furthermore, recording audio is more likely to raise privacy concerns with event attendees or event organizers. It is our experience from prior data collection efforts that the perception of audio recording, even when using more privacy-preserving low frequencies is a deterrent for participant recruitment. In fact, none of the datasets created for the study of free-standing conversations contain raw high-frequency audio [89, 95].

Instead, wearable sensors like the sociometric badge [88, 307] and mobile phone sensing [202] have been used in mingling settings to capture lower-fidelity signals that obscure the content of conversations but may still capture speaking status [308]. However, the noise introduced by the large number of sound sources in a crowd makes it challenging to distinguish speakers in a low-frequency recording [88].

For these reasons, the possibility of detecting speaking from body movement alone without access to audio offers an appealing privacy-sensitive solution to these problems. It has long been observed that hand and head gestures frequently co-occur with speech [206, 207] while being salient cues with similar motion characteristics across people.

The video modality commonly present in most in-the-wild mingling datasets [88, 89, 95] offers a convenient way to observe these gestures, with several methods having been proposed to detect speaking status from video [80, 127, 128]. So far, however, none of these methods have been trained using ground truth obtained from high-quality individual audio recordings. Video-based annotation or noisy low-frequency signals have been used instead. While video-based annotations of speaking status provide a valid supervisory signal, it is currently unclear if audio-based labels are of higher quality given that shorter turns such as back-channels are not so easily observable. For example, access to reliable back-channel annotations is an important step for social involvement detection.

In this work, we present the first dataset for studying speaking status segmentation from body movement in the wild with raw, high-quality audio from personal directional Lavalier-type microphones. REWIND offers for the first time, the possibility to scrutinize



(a) Capture from one of our cameras, including the pose tracks active in the current frame. Note that we only extracted poses for subjects who wore microphones.

(b) Crops showing individual subject pose tracks.

(c) Subject wearing the accelerometer and microphone.

Figure 5.1: Captures of our dataset and data subjects.

the relationship between speech and body movements in this setting, with audio-based ground truth that is more fine-grained (higher temporal resolution) compared to previous datasets. Beyond speaking status, the more general problem of how body movement can be used to infer phenomena observed/annotated in the audio modality opens the door for the cross-modal study of other social cues and signals like laughter and back-channeling. The REWIND dataset was already used in [309] to study laughter annotation across modalities.

In addition to audio, the dataset includes three modalities capturing body movements: video, pose, and wearable acceleration. Video recordings include top-down and side-elevated views. The latter was used to automatically obtain pose tracks for the subjects in the scene. Acceleration readings were obtained from wearable devices in a badge-like form factor worn by data subjects on the chest.

Our contributions are the following:

- We introduce the REWIND dataset, the first in-the-wild mingling dataset with high-quality raw audio, video, and acceleration; automatic pose annotations, and automatic speaking status labels.
- We present results from four body-movement-based speaking status segmentation (SSS) machine learning tasks: 1) video-based SSS, 2) acceleration-based SSS and 3) pose-based SSS, and 4) multimodal (video + acceleration + pose) SSS. In comparison with previous work, we increase the resolution of our outputs by training our models to produce a probability mask for speaking status over an input segment.
- We analyze the role of the REWIND dataset in the context of speaking status segmentation from body movement, and in the wider field of the computational study of body movements, identifying potential research directions enabled by a dataset such as REWIND.

5.2 RELATED WORK

Although generic action recognition and localization tasks have received the most attention in the literature [310], some work has been concerned specifically with speaking status detection or segmentation without access to audio [227, 311], and the challenges specific

Table 5.1: Mingling datasets with speaking status labels or related data. N: number of subjects in the dataset; STFT: short-term Fourier transform values; T-D: top-down view; S-E: side-elevated view; BBs: bounding boxes; IMU: inertial measurement unit

Dataset	N	Length	Audio and Speech Labels		Body Movement Modalities		
			Audio	VAD labels	Video	Poses	Accel.
MatchNMingle [89]	92	30 min	No	From video	T-D	No (BBs)	Yes
SALSA [88]	18	-	STFT @ 30Hz	From video	S-E	No (BBs)	Yes
ConfLab [95]	48	16 min	1250Hz	From video	T-D	Yes	Yes (IMU)
REWIND (ours)	18	90 min	44100Hz	From audio	T-D & S-E	Yes	Yes

to in-the-wild mingling settings [81, 127, 128, 312–315]. Note that the terms *detection* and *segmentation* may be used arbitrarily for models operating across a wide range of time resolutions, and we use them interchangeably while underscoring the importance of higher resolutions in supporting a wider variety of research questions. In this section we review detection and segmentation work, starting with the datasets and methods most related to our scenario. Additionally, we discuss speaking status methods developed for settings other than in-the-wild mingling, and how they fail to address key challenges specific to in-the-wild mingling.

5

5.2.1 RELATED DATASETS AND METHODS IN MINGLING SETTINGS

In the study of conversational social behavior, turn-taking patterns are a fundamental unit of analysis. Therefore, most previous mingling datasets have contained speaking status annotations and have presented estimating it as a baseline task [88, 89, 95]. These datasets contain raw modalities like video and acceleration from wearable devices. Subjects are usually localized in videos using bounding boxes, except for the recent ConfLab dataset [95] which contains full-body keypoint annotations. Table 5.1 presents an overview of existing mingling datasets used for speaking status detection. Follow-up work using some of these datasets has addressed speaking status detection from video by classifying 3-second windows as speech / non-speech [80, 127, 128, 316, 317]. Cabrera-Quiros et al. [80] presented MILES, a method using multiple instance learning to classify bags of dense trajectories calculated over 3-second windows. In a MediaEval multimedia evaluation benchmark addressing this task, Fisher Vectors were also explored as an alternative to represent dense trajectories [128]. Wang et al. showed significant improvements over both of these methods using a 3D CNN method on 1-second windows.

Poses, often derived from video frames, are another modality of interest due to their lower dimensionality compared to raw video and their ability to capture gestures. Individual pose detections have been used in action recognition work, both as standalone inputs and in combination with video, to provide precise localization information [124, 125]. However, the application of this work to speaking status detection in the wild has been limited [317]. It is unknown whether body-pose-based methods can reach the same performance as video-based ones or the effect that different pose detection approaches have on performance.

Despite their advantages, video and pose inputs to action recognition models are affected by subject occlusion, cross-contamination, poor lighting conditions, and differences in perspective, orientation, and distance to the camera. Wearable accelerometers

circumvent the aforementioned challenges and can capture subtle body movements in space. Accelerometer readings from a smart ID badge hung around the neck (Figure 5.1c) have long been studied for recognizing actions in mingling settings [85, 210]. Hung et al. [85] explored recognition of actions like gesturing, laughing, and speaking, obtaining the highest performance among them for speaking detection. The MediaEval multimedia evaluation benchmark also evaluated acceleration-based methods [316] on a subset of the MatchNMingle dataset [89], which used the same accelerometer sensors used in this work. Here, CNNs for time series have been shown to improve over traditional classifiers [128]. Particularly, a previous approach made use of transfer-learning to obtain person-specific classifiers [81]. Multiple works have found speaking status detection performance from acceleration to be higher than that of video-based methods. [80, 81, 95, 128] As with video, all of these works used 1-second or 3-second windows for classification.

Despite all these works, a key challenge remains: existing mingling datasets [88, 89, 95] do not contain high-quality audio. In [89] and [95] speaking status has been annotated from the video. In [88] speaking status was annotated from low-resolution audio and the authors noted the difficulty of distinguishing speakers in their recordings. The lack of high-quality audio in these datasets not only limits the time resolution of the speaking status annotations (and of the trained models) but may affect the correctness of the annotations. The quality of video-based speaking status annotations has not been compared with that of audio-based ground truth. Furthermore, the lack of audio makes such datasets unusable for studying other verbal phenomena, due to the impossibility of annotating them.

5.2.2 SPEAKING STATUS DETECTION IN NON-MINGLING SETTINGS

In work not specific to in-the-wild mingling settings, researchers have addressed the problem of speaking status detection/segmentation (also termed Voice Activity Detection), especially from upper-body shots of people in videos [226]. Beyan et al. [227] introduced the RealVAD dataset, consisting of a single-camera frontal recording of a panel discussion where 9 subjects take turns speaking. The RealVAD method presented in the same paper adapts CNN-extracted features from one speaker to another to improve performance. Previous work presented the Columbia dataset [318], set in a similar panel discussion setting. These datasets contain higher resolution, audio-based labels compared to mingling-specific datasets (Section 5.2.1). However, they have some fundamental differences with the mingling setting. In particular, freedom of movement in mingling datasets creates the challenge of learning from data points with a variety of camera angles, occlusion levels, orientations, and distances (of subjects respective to the camera) not present in panel discussion datasets. Perhaps due to the absence of occlusion challenges, the use of wearable acceleration has not been explored in these datasets. Furthermore, panel discussions have specific dynamics with most often a single speaker at a time. Another related task is speaker naming in movies [228, 229]. However, here it is normally assumed that the algorithm has access to audio.

5.3 DATA ACQUISITION

Investigating speaking status in a naturalistic setting involves the collection of a dataset in which social interaction occurs with as little intervention as possible. Following the design

principles outlined in [95], we collected the dataset in collaboration with organizers of a special event for a business networking group. In this section, we detail the data collection procedure and setting in which our data was collected (5.3.1) along with our sensor setup (5.3.2).

5.3.1 PARTICIPANT PROCEDURE

Most participants in the networking group met regularly and many but not all of them knew each other. Participants were informed beforehand that this particular meeting would be recorded. As they arrived at the event, attendees were approached one by one and informed again of the special circumstances of the data collection. They were then informed of the data collection process and invited to donate their data. They were free to choose which sensors to wear between microphone, accelerometer, or both; or to not participate in data collection. They were then asked to sign an Informed Consent Form. Subsequently, participants were fitted with the corresponding sensors. To enable the possibility of opting out of the video modality, all participants were informed about a clearly delimited video zone where they would be recorded by our video cameras.

After this, subjects were free to move around the room and talk as they pleased. During the first half of the event (1.5hr), however, they were at times expected to attend to a speaker, a live music performance, and participate in social games and activities. In the second half of the event (1.5hr) subjects were free to mingle without interruption, as there were no more games or activities. The room was not closed and they were free to leave at any time, after returning their sensors. During both halves of the event, most of the interaction consisted of free-standing conversation, as there was little seating available. The mood appeared friendly <https://www.overleaf.com/project/62b1a75fd2b99ab59c37dd9dand> relaxed.

When participants approached to return their sensors, we asked them to fill out an exit survey indicating their experience in the event, including rating on a scale of 1-5 their perceived level of enjoyment (4.14 ± 0.79), their likelihood of attending an event like that one again (4.21 ± 0.70) and of recommending the event to others (4.12 ± 0.72), and free-form textual feedback. The survey was associated with their sensor IDs. After the event, we sent the subjects a report of their behavior, which included information (relative to the other subjects) about their speaking time, amount of motion, and number of interaction partners.

Our complete data collection process, as outlined above, was approved by the ethics board of Delft University of Technology beforehand.

5.3.2 SENSOR SETUP

We collected the following data from consenting participants:

Audio Lavalier microphones attached to the face using Lavalier tape¹, recorded speech at 44KHz. Microphones were attached to a Sennheiser SK2000 transmitter attached near the subject's waist. Transmitters communicated wirelessly with a central receiver, which stored fully synchronized audio in real time.

¹Lavalier tape, or LAV tape is a fine tape designed to cause minimal restriction to facial muscle movements. It is transparent and is used by professional theatre productions so that the microphone remains as inconspicuous as possible. This was an important design consideration to minimize discomfort and visual distractions on the participant's face

Body acceleration A custom-made wearable tri-axial accelerometer sensor was hung around the neck and rested on the chest like a smart ID badge (Figure 5.1c), recording at 20Hz.

Video 12 overhead cameras and four side-elevated cameras were placed above and in the corners of a video zone. Every camera recorded video at a resolution of 1920x1080 and 30fps. In this work, we only make use of the four side elevated cameras.

Data was collected from the moment the participant received the sensors to the moment they returned them, which varied from about 30 minutes to more than 3 hours for some subjects who stayed after the event was officially over.

5.3.3 DATA COLLECTION DETAILS

Because some participants chose to wear only one sensor or to avoid the video zone, and because of the malfunction of some of our wearable devices, not all modalities were available for all participants. Of about 100 attendees to the event, 33 wore a microphone and 52 wore an accelerometer; while 25 wore both sensors. Most of the participants interacted within the video zone.

Our recordings included periods in which the participants were expected to listen to a speaker or a performance (Section 5.3.1). We used the videos to manually find these segments and exclude them from our experiments as they deviated from our setting of interest.

5.4 DATA ANNOTATION

In this section, we explain how the speaking status labels were generated using the individual audio recordings (Section 5.4.1) and semi-automatically for poses from side-elevated videos (5.4.2).

5.4.1 AUTOMATIC AUDIO-BASED SPEAKING STATUS ANNOTATION

Speaking status is generally labeled as a binary variable, where a positive value indicates voice activity from a target subject. The availability of high-quality audio recordings from head-worn microphones allowed us to automatically obtain labels for speaking activity via speech processing algorithms. Our task is particular in the amount of background (cocktail party) noise present and interlocutor speech present in the recordings, greater than in many speech datasets consisting of meeting or phone recordings [109, 109]. This includes interlocutor speech from subjects close to the microphone wearer, whose words could be understood from the recordings. Our goal was to ignore both interlocutor speech and background noise and obtain labels indicating the speaking status of the microphone wearer only.

We initially evaluated two VAD approaches: the *pyannote.audio* package [238, 239] and rVAD method [240]). However, we found to be unsuccessful in dealing with the noise in our audio recordings. Therefore, we decided to use a denoising step followed by a diarization step via *pyannote.audio* and *NVIDIA NeMo* libraries respectively. The complete process to generate VAD labels is as follows:

1. Loudness normalization (EBU R128) to normalize differences in audio energy across recordings (which could be due to microphone fit). We used ffmpeg’s *loudnorm* filter [319].
2. Denoising via *Speechbrain*’s SepFormer model [320] trained on the WHAM! dataset [321]. We ran the method using a 1-minute sliding window due to model input size limitations. This removed most of the cocktail party noise in the data, but not the voices of interlocutors.
3. Speaker diarization via NVIDIA NeMo [322], a pipeline including VAD, segmentation, speaker embeddings, and clustering. We found the method to effectively separate the wearer’s voice in its own cluster, distinguishing it from other speakers.
4. We manually identify the microphone wearer’s cluster in the diarization outputs, with help from video recordings to identify the speaker. We transform the speech segments into a speaking status time series using their timing information.

5

5.4.2 SEMI-AUTOMATIC POSE ANNOTATIONS

Since most pose estimation algorithms, including OpenPose, work independently on individual frames, we needed an approach to associate poses across frames. This problem has been investigated in previous work such as PoseFlow [208]. However, we found this method too computationally expensive for our use case due to the many people in the scene. We therefore implemented a semi-automatic method to obtain tracks from individual frame detections. We chose a computationally lighter method based on the observation that the chest keypoint was reliably detected and localized across frames.

Specifically, our goal was to create pose tracks by associating skeletons or poses across frames. We chose to do so by iterating over frames and assigning the poses detected in each frame to existing or new tracks. Specifically, for each frame n of the video, the pose detector outputs a set of poses given by $Q_n = \{P_{n,m} \mid m = 1, \dots, M_n\}$, where M_n is the number of people detected in frame n and $P_{n,m} = \{\mathbf{p}_{n,m,j} \mid j = 1, \dots, J\}$; $\mathbf{p}_{n,m,j} = (p_{n,m,j}^x, p_{n,m,j}^y)$ is a vector of J 2D-keypoints (or joints) representing a skeleton in the image plane. A pose track is a sequence of poses given by $U_{i,f} = \{P_{i,m}, \dots, P_{f,m}\}$, starting at frame i and ending at frame f . We associated poses in Q_n to tracks in T_n , the set of open tracks (consisting of poses from all frames up to $n - 1$). We do so by comparing poses in Q_n with the head of existing tracks whose last detected pose is not older than R_{th} frames, where R_{th} is an integer threshold parameter. In other words, we solve the assignment problem between two sets of poses: Q_n and $\{P_{f,m} \mid \forall U_{i,f} \in T_n \text{ and } n - f < R_{th}\}$. This process is repeated in order for $n = 1, \dots, N$.

The assignment problem remains to be solved. We define the distance between two poses as the Euclidean distance between their chest keypoints across frames; ie. for frames A and B , $D(P_{n_1,m_1}, P_{n_2,m_2}) = \|\mathbf{P}_{n_1,m_1}^{chest} - \mathbf{P}_{n_2,m_2}^{chest}\|$. We solve the assignment problem via the Hungarian algorithm. We add a maximum distance threshold D_{th} for assignment, such that if $D(P_{n_1,m_1}, P_{n_2,m_2}) > D_{th}$, then P_{n_1,m_1} and P_{n_2,m_2} cannot be assigned to each other. Assigned keypoints are added to the corresponding track and unassigned keypoints are assigned to a new track. When a new pose in frame n_1 is matched to a pose in a frame $n_2 \neq n_1 - 1$ (not the immediately preceding frame), we imputed the keypoints via linear interpolation to maintain the continuity of the track.

Parameters R_{th} and D_{th} were set based on a qualitative evaluation of the algorithm on a subset of the tracks. For R_{th} , we found a one-second threshold to work best creating consistency across frames without introducing significant errors. This approach resulted in high-quality tracks, with only sporadic track misassignments due to subjects walking in front of one another.

Because our goal was to obtain high-quality tracks to be able to reliably test our recognition method (see next sections), we manually inspected the dataset for track switches and corrected them by splitting the tracks. Finally, we assigned tracks to subject IDs to be able to associate with the personal acceleration readings.

5.5 DATASET STATISTICS

Due to our mixed-consent data collection design, our final dataset contained subjects with different body movement signals available. Subjects were also in the scene for varying amounts of time, and some engaged more than others in conversations. Figure 5.2 plots the speaking times per subject, calculated as the summation of all the speaking segments output by VAD (Section 5.4.1), together with an indication of the modalities available for each subject. The majority of subjects have complete information. There is quite some variation in speaking time with no clear correlation between modality and amount of time spent speaking.

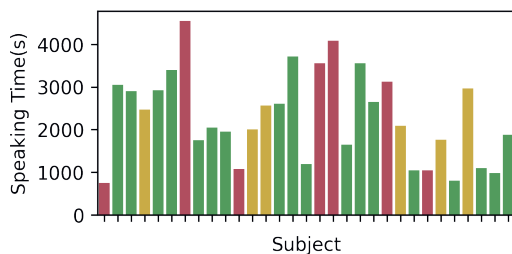


Figure 5.2: Seconds of speaking time per subject with speech data in REWIND dataset. Columns in green indicate subjects with complete information (audio, video, acceleration). Columns in yellow indicate subjects with audio and acceleration information, but who are not visible in the videos (no pose). Columns in red indicate subjects without body movement data (only audio).

To showcase the value of our automatic VAD annotations obtained from audio compared to VAD annotations in previous datasets, we compared the distribution of speaking segments in REWIND to that of the MatchNMingle dataset [312]. Figure 5.3 shows the length distribution of segments labeled as speech in both datasets. Although these speaking segments do not constitute turn-lengths, our data contains more short speech segments. Inspection of the dataset revealed these often correspond to back-channels and short utterances. We interpret that many such utterances were probably missed in previous datasets due to the use of video for speaking status annotation [95, 312].

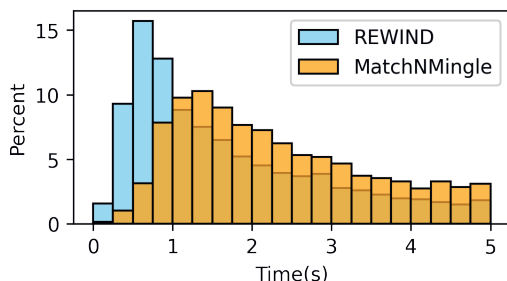


Figure 5.3: Distribution of length of contiguous speaking segments (s) in the speaking status detection labels for REWIND and MatchNMingle [80] datasets. REWIND shows greater temporal granularity (shorter segments), thanks to annotations having been obtained from audio.

5.6 BASELINES FOR AUTOMATED SPEAKING STATUS SEGMENTATION

5

In this section, we present three baseline tasks for speaking status segmentation from body motion. We start by detailing our evaluation setup, including the process for generating training and testing examples, evaluation metrics, and hyperparameter tuning. We first present the video baseline, where we train on video patches around target subjects. Second, we present a pose-based approach, where we train on pose track segments. Finally, we present a method based on acceleration readings.

5.6.1 EVALUATION SETUP

We formulated the problem as speaking status segmentation to take advantage of the high time resolution of our labels. We defined data samples as 3-second segments of behavior, following previous work which found this length to be close to ideal for speaking status detection [81, 312, 314, 317]. However, unlike prior works which would predict a single label per 3-second window, our method provides finer granularity by predicting the entire binary speaking status time series (20Hz) for a given data sample.

For consistency across models, we excluded subjects not containing all three modalities (video, pose, acceleration). We leave it to future studies to investigate further the trade-offs. This left us with a total of 18 subjects. We obtained 3-second windows by splitting pose tracks using a sliding window with no overlap. This resulted in a dataset with 16403 examples in total.

For all network hyper-parameters, we used their default values, which were found to work best over a variety of datasets. For setting the number of training epochs, we used a held-out set of 10% of the dataset as a validation set. Here, due to the small number of subjects, we partitioned by data point rather than by training subject. Using the rest of the dataset (90%) we evaluated via 3-fold cross-validation at the subject level, to measure generalization to new subjects. We measured performance via Area under the ROC curve (AUC), where we treated every window element (of which there were 60) as one separate prediction. Although metrics like Intersection over Union (IoU) are often used in segmentation, we made use of AUC due to its robustness against class imbalance and

because it is the most conservative estimate for the localization of speaking samples.

5.6.2 VIDEO-BASED SPEAKING STATUS SEGMENTATION

Due to the relatively small size of our dataset, training state-of-the-art video action recognition methods from scratch would be infeasible. We focused on approaches with pre-trained models available to use as feature extractors. Among those, 3D convolutional neural networks (CNNs) are known to reliably achieve top performances in action recognition benchmarks. We decided to make use of a 3D ResNet pretrained on Kinetics-400, a large action recognition dataset with 400 action classes and over 300000 labeled video clips. The network implementation and models are available as part of the *Pytorchvideo* library [323]. To adapt the network to our outputs, we implemented a custom network head to apply pooling and convolution operations over the spatial and channel dimensions and up-sample the time dimension to the length of the target mask (60). Details are in Appendix 5.A.

5.6.3 POSE AND ACCELERATION-BASED SPEAKING STATUS SETECTORS

For pose and acceleration, we made use of a ResNet variant for time series, implemented as part of the *tsai* library [306]. Given the much lower dimensionality of these modalities (when compared to video) we trained both models from scratch. As with the video method, we implemented segmentation network heads to output masks of length 60. Details of the network heads are in Appendix 5.A.

5.6.4 MULTIMODAL SPEAKING STATUS SEGMENTATION

Given that our hypothesis about the link between speech and body movement is modality agnostic, we included a baseline combining all three forms of body movement representation. The assumption is that the granularity of poses will allow the method, for example, to distinguish head and hand gestures from general body movement. The videos will capture more subtle movements (that might be obscured by noise from associating the detected pose skeletons between frames) and shape characteristics of a person's behavior. Finally, the acceleration may capture sub-pixel and 3D characteristics of the movements that are not discernible in the video. For models using a combination of video, poses, and acceleration inputs, we merged the architectures above by averaging their output masks (output fusion). Network heads were retrained from scratch.

The results of our evaluation are presented in table 5.2. Results suggest the superiority of combining modalities for the task. For video and acceleration, these results align with previous work on speaking status detection which found these modalities to perform comparably [80, 127]. However, our pose method performed poorly compared to video and acceleration methods. A likely cause is the noisy nature of the poses. While our approach delivered reasonable track association performance, the fact that tracks are extracted independently per frame introduced significant noise across frames. The relative nature of poses would likely make it harder for the model to separate speech-related gestures from pose noise. It is also possible that using pre-training in a state-of-the-art skeleton action recognition method could improve these results. Note, however, that large pre-trained skeleton action recognition methods are often trained on sequences with more than one

Table 5.2: Results of our baseline evaluations.

Method	AUC
Video 3D-CNN	0.615
Pose CNN	0.530
Accel. CNN	0.634
Video + Pose + Accel	0.648

skeleton, and do not use the same skeleton definition (input size and semantics) [288]. This makes adapting them to our problem not trivial.

5.7 DISCUSSION AND CONCLUSIONS

With REWIND, we contribute the first dataset for speaking status segmentation recorded in a real-life mingling scenario and with high-quality individual audio recordings, and derived speaking status annotations. Although the use case of the dataset showcased in this paper is speaking status segmentation, REWIND creates opportunities for research beyond this task, in the automatic detection/segmentation of body movement manifestations of social signals in general. In this section we discuss the implications of the dataset, starting with the results presented in this paper, and following with a discussion of the possibilities brought about by REWIND in other related tasks.

REWIND AS A DATASET FOR NO-AUDIO SPEAKING STATUS SEGMENTATION.

Previous speaking status works have shown that it is possible to perform speech/non-speech classification from body movement information [81, 89, 95], as well as classifying the current speaker in a group [88]. The same lack of audio makes it impossible to verify this, but our results comparing turn length distributions of REWIND with MatchNMingle, suggest that many short ($< 1s$) speech segments are either lost or aggregated into longer segments (losing granularity). Furthermore, annotating speech from visual modalities may mean that speech segments were not missed at random, but that the most visually subtle speech is missed, resulting in an undesirable bias. A more comprehensive study analogous to [309] would shed more light on this.

With REWIND, we have presented an approach to record, process, and automatically extract speaking status labels from a small mingling crowd. This approach resulted in a different distribution of speech segment lengths when compared to video-based labels due to a shift towards shorter segments (higher granularity). Furthermore, the time resolution of our labels enables a task not previously attempted: segmentation of speaking status. The availability of ground truth audio means that our annotations are easy to verify manually and to further refine automatically in the future.

One drawback is that the guaranteed level of synchronization of the REWIND dataset is within latencies of 1s. More precise estimations would require collecting new data with high-quality audio using a similar multi-sensor synchronization strategy as [95].

REWIND AND THE STUDY OF BODY MOVEMENT.

REWIND, through its high-quality audio recordings, creates opportunities for studying the relationship between vocal production and body movement in naturalistic social interaction. Social actions with a vocal component such as laughter and back-channeling have been studied in the past in relation to body movement [107, 309, 324]. Higher-level multimodal constructs such as affect, enjoyment, or engagement also have manifestations in both vocal production and body movement. With REWIND, in addition to providing aggregate self-reports of enjoyment, we provide the raw data necessary for performing third-party annotations of these constructs from audio, video, or audiovisual information at a higher temporal resolution.

This creates opportunities for using REWIND to train action detectors using different input and labeling modalities. Furthermore, it also allows for exploring the effect that different labeling conditions (eg. video-based labeling) have on both label reliability and model performance. This can further our understanding of the trade-offs in labeling inherently multimodal phenomena from limited modalities such as video and audio. A study using the REWIND dataset has already addressed such questions in the context of laughter detection, intensity estimation, and segmentation [309].

5.7.1 EFFICACY OF POSE-BASED ANALYSIS

One limitation of REWIND lies in the quality of the pose tracks. Due to challenges like occlusion and cross-contamination, pose tracks obtained from our system are noisy and may miss subjects, especially those far away from the camera. While we consider our tracks to be enough for many applications including evaluation of action recognition methods, they may not be enough for evaluating tasks like person detection or tracking, where the goal is to detect/track all the people in the frame.

MIXED-MODALITY CONSENT: LIMITATION AND OPPORTUNITY

Finally, the reader may consider that another limitation of REWIND is that many users in the scene did not wear our instruments (Section 5.5). This means that analysis or prediction of social signals from group-level information is difficult with this dataset. An exception is that video data is available for entire groups, and could be used to predict individual variables (eg. speaking status). While this can be seen as a limitation, it gives us the opportunity to investigate such mixed consent settings and the analysis of partially complete data.

ACKNOWLEDGEMENTS

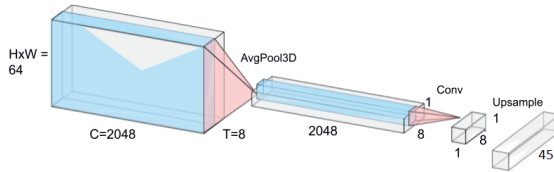
This research is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606. We acknowledge Bernd Dudzik, Xianhao Ni, Xiang Teng and Alessio Rosatelli for their assistance during the data collection event.

Appendices

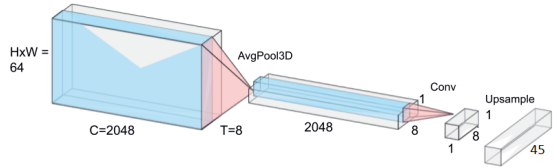
REWIND dataset: privacy-preserving speaking status segmentation from multimodal body movement signals in the wild

5.A NETWORK DETAILS

Figure 5.4 presents the architecture of the segmentation heads. For all models, we apply pooling and convolution operations over the spatial and channel dimensions and up-sample the time dimension to the length of the target segmentation mask (60). Output masks are averaged for multimodal methods.



(a) Time series ResNet head (acceleration modality).



(b) Video ResNet (slow model) head.

Figure 5.4: Segmentation heads for acceleration and video models. The first block represents the feature map before the head of the ResNet model, for each modality method. Subsequent operations pool and convolve over the spatial and channel dimensions, and up-sample the time dimension.

6

COVLEE: AN EXTENSIBLE WEB FRAMEWORK FOR CONTINUOUS-TIME ANNOTATION OF HUMAN BEHAVIOR

6

Continuous-time annotation, where subjects annotate data while watching the continuous media (video, audio, or time series in general) has traditionally been applied to the annotation of continuous-value variables like arousal and valence in Affective Computing. On the other hand, machine perception tasks are most often annotated using frame-wise techniques. For actions, annotators find the start and end frame of the action of interest using a graphical interface. However, given the duration of the videos generally annotated in social interaction datasets, this can be a slow and frustrating process. It usually involves pausing the video at the onset or offset of the action and scrolling back and forth to identify the precise moment. A continuous annotation system, where annotators are asked to press a key when they perceive the target action, can reduce the time necessary to annotate, especially when single subjects are annotated for long periods. Keypoint annotations, where the task is to follow a particular point of interest in a video (e.g., a body joint) may also be done continuously. We present the Covlee web framework, a software package designed to support online continuous annotation tasks, with crowd-sourcing capabilities. We present results from case studies of continuous annotation of body poses (keypoints) and speaking (action) on an in-the-wild social interaction dataset. In the case of keypoints, we present a new technique for easily following a keypoint in a video using the mouse cursor. We found this technique to significantly reduce annotation times with no adverse effect on inter-annotator agreement. For action annotation, we used continuous annotation techniques to obtain binary speaking status labels and annotator ratings of confidence on those labels. Covlee is free software, available as a Python package documented at josedvq.github.io/covlee.

6.1 INTRODUCTION

Annotating human behavior for machine perception tasks involves extracting fine-grained facial and body behavior. Depending on the tasks or research questions being investigated, annotations may, for example, look to describe the movement and spatial location of a person via bounding box or keypoint annotations, indicate what actions are being performed by such person via binary action annotations, or describe the state of the person by annotating constructs like enjoyment or involvement.

Clearly, not all machine perception tasks and associated annotation tasks are created equal. Importantly, datasets containing human behavior vary widely in the number of subjects present in the dataset and the length of time each subject is recorded.

For example, most benchmarks in computer vision tasks of action recognition and pose estimation use still images or short video clips for training and benchmarking [220]. This often means many data subjects in different environments, each recorded (and annotated) for a short period. This is desirable when the goal is to maximize data diversity to enhance the system's robustness. In these tasks, annotations for keypoints are performed on individual frames, and videos are labeled with a single action.

In contrast, in applied machine learning within the social signal processing research, interacting subjects in audiovisual datasets need to be tracked and annotated for periods ranging from a few minutes to several hours [88, 89, 109] which is necessary to capture and study social interaction dynamics. Similarly in the affective computing community, datasets often involve annotating interactions lasting one hour or longer [325]. Behavior analysis datasets often have fewer data subjects, recorded for longer periods. Other applied fields working with in-the-wild data like surveillance and sports action recognition often require tracking subjects for long periods [326].

The annotation challenge is compounded when datasets are acquired in the wild (without the benefits of lab-based, highly instrumented recording spaces), meaning that automatic techniques for subject detection, tracking, and pose estimation are not applicable. Obtaining the same level of detail of human behavior in these settings is often prohibitive in terms of the manual labor or equivalently financial cost involved.

A second key characteristic of many human behavior annotations, especially those of actions and higher-order constructs, is the central role of temporal context in perception. While *simple* tasks such as the annotation of body joints in a video can be considered free of temporal context (ie. a single frame can be meaningfully annotated), annotating concepts which require a judgment about intention, such as *the use of sarcasm* or *dominant laughter* requires a judgment that can only be done with access to temporal context (ie. the past) of the interaction.

Annotating human subjects for long periods while having access to temporal context has created a need for annotation tooling that we argue is not covered by existing annotation tools and techniques. In this chapter, we present a software framework offering a technical solution to this problem.

Continuous-time annotation refers to annotations being carried out in real time while the target media is being watched without pause. Traditionally, continuous-time annotation has almost exclusively been applied to the annotation of affect of a target subject, usually being observed in a video. Affect has been annotated via the variables in the circumplex model of affect: arousal and valence (considered continuous variables). Joint annotation of

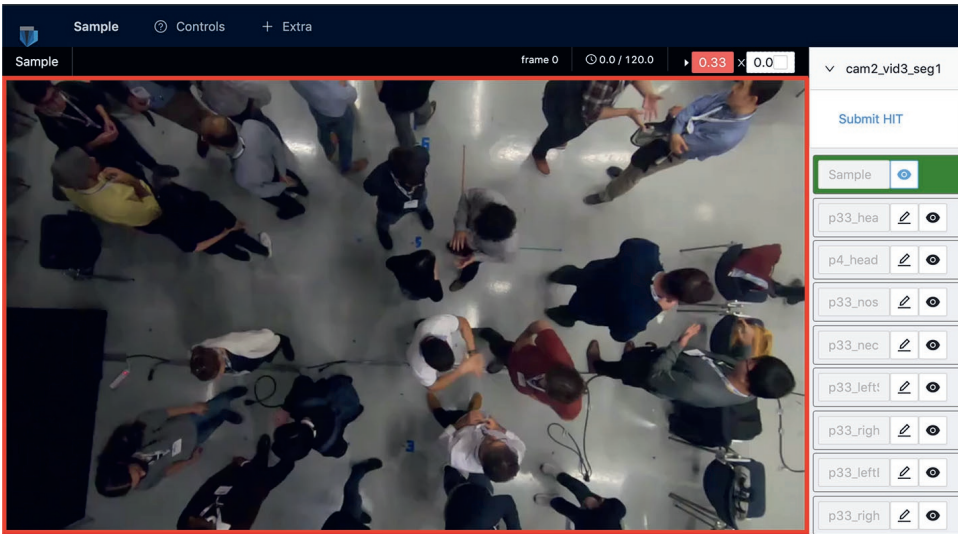


Figure 6.1: The Covfee keypoints annotation interface.

both variables was first proposed, where the annotator controls the position of a cursor within a labeled diagram (2D annotation) using their mouse [114]. Further developments split the annotation process into the separate annotation of arousal and valence [327]. Since then, continuous-time annotation has been used to annotate multiple datasets [328, 329], more modern tools have been developed [115, 116, 330], and the best way to make use of continuous-value annotations taking into account human biases has been researched [111, 113]; all within the context of affective computing.

In this chapter, however, we treat continuous-time annotation as a general technique applicable to different types of variables and target media. In addition to continuous affect annotation, examples of continuous-time annotation include holding down a keyboard key to indicate that a person in a video is speaking, following a person's hand with the mouse cursor to indicate its position, or controlling a continuous slider using the mouse to rate the perceived level of engagement of a person in an interaction.

We investigate the power of continuous-time annotation to improve annotation times when labeling human subjects for long periods. Its continuous nature has the advantage of facilitating the perception of temporal context, potentially improving the quality of annotations through better annotator judgment of the target action or construct.

Annotating subjects for long periods can be made more feasible by leveraging crowd-sourcing, splitting the load among multiple annotators. In crowd-sourcing, remote workers are paid to perform HITs (human intelligence tasks) consisting of units of work to be completed by one annotator (usually taking a few minutes to complete). Given its use in different fields, notably computer vision, and human behavior analysis fields, we leveraged its benefits by giving our continuous annotation framework crowd-sourcing capabilities.

Covfee, brings together the possibilities of continuous annotation with those of crowd-sourcing into a web-based annotation framework. Because annotation techniques are

often task-specific and continuous annotation is a nascent field in need of experimentation, we designed and documented Covlee as an extensible framework, to encourage users to implement new techniques with as little effort as possible. Continuous annotation interfaces for affect, action annotation and keypoint annotation are applications of the framework.

Our contributions are the following:

- We present Covlee, an open-source web annotation framework with crowd-sourcing support, implementing continuous action and keypoint annotation out-of-the-box. Covlee supports the implementation of custom continuous annotation tasks with different media types and user interfaces taking advantage of existing capabilities for data serialization/storage, crowd-sourcing, qualification testing, and annotation tracking and monitoring. Annotation tasks to be implemented in Covlee may range from existing techniques for rating continuous variables like affective dimensions [111] to novel techniques for vision tasks such as the ones presented in this paper.
- We present a case study involving the use of Covlee for the annotation of human body joints in an in-the-wild dataset. We present comparative results against a traditional annotation method and found a nearly three-fold improvement in annotation time with no loss in inter-annotator agreement.
- We present a second application of Covlee for the efficient annotation of actions in the same social interaction dataset. We analyze annotations of speaking status via a continuous binary interface and confidence ratings for those annotations.
- We discuss the advantages and disadvantages of applying continuous annotation to human behavior datasets. Based on both case studies, we provide recommendations and a discussion of other potential use cases for Covlee.

The remainder of the chapter is organized as follows. In section 6.2 we start with a summary of related work and its relevance to the Covlee framework and its tasks. In section 6.3 we present the Covlee framework, summarize our main design requirements and decisions, and present its main features for both basic users looking to annotate data or advanced users looking to implement new tasks using Covlee. In section 6.4 we present two case studies using the Covlee framework for new types of continuous annotation: the fully manual annotation of human body joints using the mouse as a tracking device; and the binary annotation of actions in a social scene. We end by discussing and reflecting on these case studies and the role of continuous techniques in human behavior annotation in Section 6.5.

6.2 RELATED WORK

In this section, we start by reviewing work on manual annotation tools, with a focus on web computer vision and time series annotation tools. We go on to review work specific to continuous annotation, most of which relates to the annotation of human subjects.

6.2.1 MANUALLY ANNOTATING KEYPOINTS AND ACTIONS

Of particular importance in computer vision tasks involving human data subjects are the tasks of pose estimation [97, 257], or keypoint estimation in general, and the task of action recognition or localization [220], both of which we address in this chapter.

Keypoint annotation involves the labeling of important points in an object of interest. This could be hand joints, facial landmarks, or object keypoints. Keypoint annotation is supported in tools like Vatic and CVAT via image-level annotations, performed every N frames, and interpolated in between. This is however a time-consuming process whose accuracy is limited by the interpolation step, particularly if a keypoint being tracked moves with highly varying levels of acceleration. The number of frames to skip should be few enough to avoid under-fitting a particular trajectory of a keypoint whilst still being large enough to minimize manual effort. It is also unclear how to deal with frequent occlusion of the target keypoint in this scenario and annotating for such occlusion makes the process slower.

As a result of these challenges, many works involving the tracking of many individuals in a social scene have reverted to using bounding boxes for subject localization despite the fact that full-body poses contain richer information [89]. Others have reverted to using a much smaller set of skeletal points such as just using head positions and orientations [88]. Parallel to this, there is a growing community working on the detection of actions directly from skeletal data [284] given the emergence of large-scale keypoint data that has been automatically generated in highly instrumented lab environments [257]. Being able to annotate body keypoints in in-the-wild settings provides a sound basis for researchers in these areas to transition to working on more realistic natural settings.

The first step in action annotation involves the localization of actions of interest in a recording. In social interaction datasets such recordings often capture a large social event [89], multiple meetings [109] or conversations, spanning dozens of hours of individual interaction and requiring a time-consuming effort to annotate. To this end, actions are traditionally localized using a mouse and graphical interface. In tools such as ELAN [120], the user localizes the start and end frame of the action, which is then annotated by drawing an interval in a timeline. In tools such as Vatic and CVAT, actions are annotated via flags, which are turned on for the frame when the action is deemed to start, and off for the end of the action. Both of these approaches require the user to pause the video every time an action is recognized. This has the drawback of slowing down the process and making it harder for the annotators to follow the flow or dynamics of the interaction, or media in general.

An important consideration when annotating body keypoints is the annotation of occlusion: when the target body joint is not visible, due to being occluded by another object/person in the scene, or possibly the same person (self-occlusion). Rather than being constant, in many in-the-wild datasets, body keypoints may become visible and occluded frequently when bodies gesture, change posture, or move around in the scene. Occlusion signals are important for training pose estimation methods, and are included in several datasets for pose estimation [231, 283]. Networks designed to learn from the occlusion signals have been shown to improve performance on pose estimation image datasets [331, 332].

6.2.2 CROWDSOURCING ANNOTATIONS

The advent of deep learning in computer vision has resulted in algorithms requiring large amounts of data to reach state-of-the-art performance. In video-based tasks like action classification (recognition) and localization, this has required the labeling of datasets with hundreds of hours of video, used now as benchmarking datasets. Improvements in the related task of pose estimation [97–99], commonly trained from images, have been possible thanks to image datasets with tens of thousands of examples [231, 283]. This process has come together with the development of annotation tools capable of supporting at least a range of canonical tasks: keypoint annotation, bounding box annotation, image segmentation, and temporal (action) labeling.

With the move towards online services, crowd-sourcing annotations has gained relevance in computer vision. This trend led to the collection and annotation of large datasets completely online [222]. In human behavior analysis in particular, crowd-sourcing has been used to annotate actions and person bounding boxes within the social signal processing community [89] and more extensively used in the affective computing community [333–335], where techniques for improving reliability in the crowd-sourcing setting have been explored [336].

In a comprehensive paper on the subject of crowd-sourcing Vondrick et al. [258] use the Vatic tool to provide a series of insights related to online video annotations. Even though annotators used traditional frame-level annotation techniques, some of their insights are relevant to annotation processes and tools in general, and we summarize them here. When annotating body joints, they found that annotators are more efficient and prefer to annotate one joint at a time throughout the whole video compared to annotating one image (all joints) at a time. Another important observation was that annotators “rely on the motion of objects to correctly decode the scene”, and that “the user must watch the video play to correctly track [an object]” [258, p.7]. Both of these are default choices in the continuous annotation paradigm, where the annotation technique must be simplified to be done while the video plays.

In the same paper, authors concluded that larger tasks, where a single annotator annotates all objects in a video are better than smaller tasks, such as different annotators annotating single objects. This is likely due to the overhead involved in familiarizing oneself with the scene to annotate. They also found that a constrained interface without too many choices will result in better annotation times, compared to more flexible ones. The authors address the importance of filtering workers through qualification tasks, stating that “because video annotation is hard, we found that most workers, despite accepting the task, do not have the necessary patience or skill to be accurate annotators.” We take advantage of these important insights in the design of the Covfee framework and associated annotation techniques.

6.2.3 CONTINUOUS-TIME ANNOTATION

The term *continuous annotation* generally refers to *continuous-time* annotation. Although a precise delineation of what constitutes continuous-time annotation is not present in the literature, we will treat it as an umbrella term that describes the process of performing an annotation task while the target media is being watched (possibly in real time), usually without any pauses. A distinction must be made from continuous-value annotation which

refers to the annotation of continuous variables in general which could also be carried out as a post-hoc annotation step. Although mostly applied to audiovisual recordings, continuous annotation is not limited to this set of modalities and may apply to any sensory experience such as listening to an audio recording or watching a live performance.

Continuous-time annotation started with the continuous recording of emotional states with Feeltrace, an instrument designed to let observers *track the emotional content of a stimulus as they perceive it over time* [114, p.1]. The interface consists of a circle, with dimensions corresponding roughly to arousal and valence [337], the dimensions in the widely-used circumplex model of affect [337, 338]. This type of continuous annotation allowed observers to describe an emotional state by moving a pointer within the circle using their mouse. The newer GTrace technique [327], presented as a successor to Feeltrace supported one-dimensional annotations of valence and arousal with visual feedback markers on a desktop application.

Continuous annotation has since been used in the affective computing community for the annotation of datasets for affect through GTrace-type interfaces. In datasets like DEAP [339], SEMAINE [32], RECOLA [271] and DECAF [340], valence and arousal were annotated separately using a mouse-controlled graphical interface. More recently, datasets like SEWA [328] and CASE [329] have moved to the use of joysticks for simultaneous annotation of arousal and valence. The reasons cited by Sharma et al. [329] are that separate annotation of arousal and valence does not account for the relationship between them, and that “mouse-based annotation tools are generally less ergonomic than joysticks”. However, Metallinou and Narayanan [341] more precisely state that Feeltrace and GTrace require continuously pressing the mouse to annotate, which is tiring when annotating long videos. CARMA [330] and DARMA [115] are other desktop tools for continuous affect annotation with mouse and joysticks respectively.

More recently, RankTrace [111] addressed the problem that humans are bad at maintaining references of continuous values, which is supported by theories such as the adaptation level theory. This theory suggests that “humans cannot maintain a constant value about subjective notions; instead, their preferences are made on a pairwise comparison basis using an internal ordinal scale” [111, p.1]. Their interface instead captures unbounded annotations, which are then interpreted using their gradient. They showed that the gradient of the unbounded annotations was a better predictor of skin conductance (as a correlate of emotion) than the absolute value of the annotations. They performed annotations using a hardware wheel for input. The issue of interpreting continuous-valued annotations directly relates to the question of how to measure agreement between annotators. To this end, Booth and Narayanan [113] designed an ordinal agreement measure for continuous-time, continuous-value annotations, based on the observation that annotators approximately preserve rank ordering and capture trends (increasing or decreasing) when annotating continuous values. These findings may limit the utility of continuous-value annotations (since they cannot be reliably compared absolutely). It is however unclear to what extent they generalize to the annotation of less subjective variables.

A major drawback of the previously-mentioned tools is that they were only implemented as Windows applications, making them unusable in a crowd-sourcing setting, and therefore hard to scale for use in large datasets. Web-based applications, in contrast, offer a lower barrier to access for annotators, do not require the annotators to store a

local copy of the annotated media, and may support the crowd-sourcing of annotations in online marketplaces. The data storage issue is an important one when the data to be annotated is considered privacy sensitive. Streaming data for annotation through a web interface mitigates intentional or unintentional data privacy violations such as forgetting to delete the raw data after it has been used for annotation. PAGAN [116] is possibly the first web-based tool for continuous annotation, with support for GTrace and RankTrace, as well as binary annotations. PAGAN specializes in affect annotations and is not geared toward supporting the implementation of custom techniques.

Continuous-time annotation has some inherent delay due to the time that annotators take to react and process their perception. This could potentially impair the performance of systems that are developed to learn from such data. Some recent efforts have concentrated on the study of these delays and how they can be corrected in the context of affect annotation [118]. Mariooryad and Busso [117] align annotations by maximizing the mutual information between annotations and expressive behaviors as captured by facial action units and speech features. Khorram et al. [119] presents a convolutional network capable of jointly aligning and predicting continuous emotion annotations via a time-shifted low-pass filter. Although these works show improvements in regression performance with respect to baselines without correction, it is unclear how much such correction methods can improve performance, as the true delays in the studied datasets are unknown. Furthermore, Mariooryad and Busso [117] found no significant differences when using a constant delay, compared to their data-driven approach. Although work on delays has been exclusively done in the context of continuous-valued affect annotations, and delays in annotation are likely task-specific due to different stimuli processing times, some degree of human delay is inherent to all kinds of continuous annotation. It seems pertinent for researchers developing machine perception systems trained with continuously annotated data to be mindful of the potential effects (if any) of delay.

6

In summary, annotation in computer vision and continuous annotation are two completely disjoint fields in the literature. The former has focused on image-level techniques aided by interpolation for the annotation of keypoints for pose estimation tags, and the use of binary flags for the annotation of actions, used in action localization tasks. It is however unknown how such annotation techniques compare to continuous ones in time efficiency and annotation quality, and to what extent their non-continuous nature affects annotations heavily dependent on temporal context. A reason for this is that continuous annotation literature has almost exclusively focused on the subject of affect, which has resulted in very specific techniques, tools, and insights for continuous-time annotation of continuous affect variables. This means there is little study of the phenomenon of continuous annotation in its more general form, which may involve different modalities, input devices, and interfaces. The lack of software support for the implementation of continuous annotation tasks also limits the broader study of this topic. To this end, we hope that Covfee lowers the entry level for more researchers to explore continuous annotation in more settings allowing for a better understanding of its potential.

6.3 THE COVFFEE FRAMEWORK FOR CONTINUOUS ANNOTATION

Covfee was born out of the need for an advanced framework for both annotating existing datasets and researching questions related to continuous annotation. The target user of Covfee is thus a researcher aiming to annotate a human behavior dataset or to use or implement novel continuous annotation interfaces for research purposes. As such, Covfee was built with a set of main broad requirements:

- To be under an open source license and documented online. Covfee has been released under an MIT license, a permissive license enabling among others the copying, modification, and redistribution of the software without limitations.
- To be easy to install, and launchable on a local web browser from a command line.
- To be deployable in a public server for online annotation. This is also necessary to support crowd-sourcing annotations in online marketplaces.
- To support large annotation processes consisting of hundreds of HITs, with each ranging from seconds to hours in length (of the target media).
- To implement client-server communication of annotations and storage on the server. Annotations should be buffered (to prevent data loss from network errors) and submitted to the server where Covfee is deployed, where they should be easy to download by the requester.
- That annotation techniques implemented in Covfee (eg. binary annotation of videos, video keypoint annotation) are easy to reuse and re-deploy.
- To support additional functionality that is useful in an online annotation process: requesting non-continuous feedback from annotators (eg. demographics, experience feedback, etc), requiring agreement to terms and conditions (eg. an EULA) before getting access to the data, and providing rich annotation instructions (images, videos, tooltips) for users.
- To support automated qualification tasks via the implementation of a validation method in Python. Validation methods receive the annotations and return a boolean decision on whether the submission passes the qualification test or not.
- That new custom tasks can be implemented with a basic knowledge of Javascript / Typescript by implementing a class with a specific interface; much like writing a custom network in modern deep learning frameworks can be done by implementing methods of a subclass
- To run in most modern desktop browsers. We do not discard making Covfee tasks usable in mobile devices in the future/ However, due to the additional implementation effort this would require, we decided to start with desktop browser support only. Note that since tasks in Covfee may be custom and use any browser features or APIs, particular tasks may have more reduced compatibility than Covfee as a whole.

Covfee should provide a way to test for browser compatibility and instruct annotators to use a compatible browser before they start working on a task.

To support these broad goals, we implemented Covfee as a Python package available in the Python Package Index. Once installed, Covfee's administration panel can be started in a browser from the command line. The main building blocks of Covfee are shown in Figure 6.2. The web application was implemented in Typescript as a one-page application in the popular React web framework. The web server makes use of the Flask framework and a SQLite database for annotation storage.

6.3.1 THE COVLEE SPECIFICATION FILE

An important question in the design of Covfee was how to let requestors describe the HITs to be created in Covfee (ie. how would a researcher use Covfee?). Existing online annotation tools let the requester create HITs using a graphical interface where media files can be uploaded and the variables to be annotated are specified. Each HIT maps to an interface with tools to support different annotation techniques (eg. drawing bounding boxes, keypoints, and setting binary flags). An annotator is expected to navigate this rich interface to annotate the requested variables. Designing Covfee in this way would have several major drawbacks. First, for large annotation processes with hundreds of files to be annotated (each of which would map to a different HIT), specifying HITs using a graphical interface would be cumbersome for the requester. Second, having a single rich interface with tools and options for different annotation techniques is not desirable. Richer interfaces with many options were found by Vondrick et al. [258] to lead to information overload for annotators. Ideally, the annotation interface should only contain the tools and information necessary to complete the HIT.

To avoid these drawbacks, we designed Covfee to read a JSON (Javascript Object Notation) specification file describing the HITs to be created as input. Instead of uploading media files using a graphical interface, URLs to the media files to annotate are part of the specification. Using a file following a particular structure makes it easy for the requester to generate this file using the programming language of their preference, an advantage for large annotation projects. For smaller annotation projects the file can also be created by hand based on the examples in Covfee's documentation. Because the Covfee specification maps directly to a set of HITs, it also serves as a shareable record to help other researchers reproduce a particular annotation project.

The specification file follows a particular structure, which among other things includes:

- Project details, including the name and details of the contact person. This information is shown to annotators in case they run into an issue during the annotation process.
- A list of HITs forming part of the project. Each HIT can be reproduced multiple times via a `repeat` parameter in the specification, or using the Covfee interface. Every instance of a HIT is mapped to a URL, meant to be visited by one annotator. A HIT in Covfee consists of a set of sub-tasks, of possibly different types. For example, a HIT may contain a keypoint annotation task and a binary action annotation task.
- For each HIT, a list of tasks comprising it. The specification of a task is different depending on its parameters. For example, the specification of a keypoint annotation

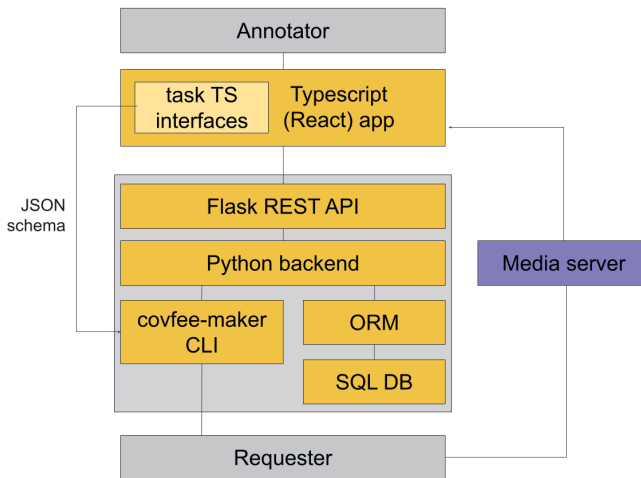


Figure 6.2: The architecture of the Covfee framework. A Python server with a relational database (SQL DB) and ORM layer (a layer mapping database tables into software objects) takes care of data storage and communication with the Typescript web application. The requester (researcher) interacts with Covfee via the *covfee make* CLI, which validates a user-provided specification of the HITs to be created. Typescript interfaces translated into JSON schema (a language for describing the structure of objects) are used as templates to validate the specification and provide friendly errors to the requester in case of mistakes in the specification.

task is different from that of an action annotation task. Each task in the specification maps to an annotation interface that is specific to its task type (ie. annotation technique). This minimizes information overload for annotators by giving them only the tools, options, and instructions relevant to the task at hand.

An example of a Covfee specification file is shown in appendix 6.A. Specification files are validated by Covfee to ensure that they have the correct structure and valid property names and values. Friendly error messages are returned indicating the location and cause of any error within the structure. This makes it easy for the user to debug their specification and avoids potentially hard-to-trace errors due to mistakes in the specification. Appendix 6.A shows an example of validation output from Covfee.

On the technical side, validation naturally requires a model or schema of what the specification should look like. The use of Typescript for the implementation of tasks in Covfee provides a natural way to do this. Typescript interfaces are used to specify the shape and parameters of each task's specification. Covfee internally translates these interfaces into JSON Schema [342], a vocabulary for the validation of JSON documents. These JSON schemas are used by the covfee CLI to validate the JSON structures. Figure 6.2 shows a diagram of Covfee's architecture, including this process.

6.3.2 ONLINE WORKFLOW

Figure 6.3 diagrams the workflow in Covfee. The main participants are the requester (researcher) and the annotators. Annotators get access to a Covfee interface generated by the requester using the framework.

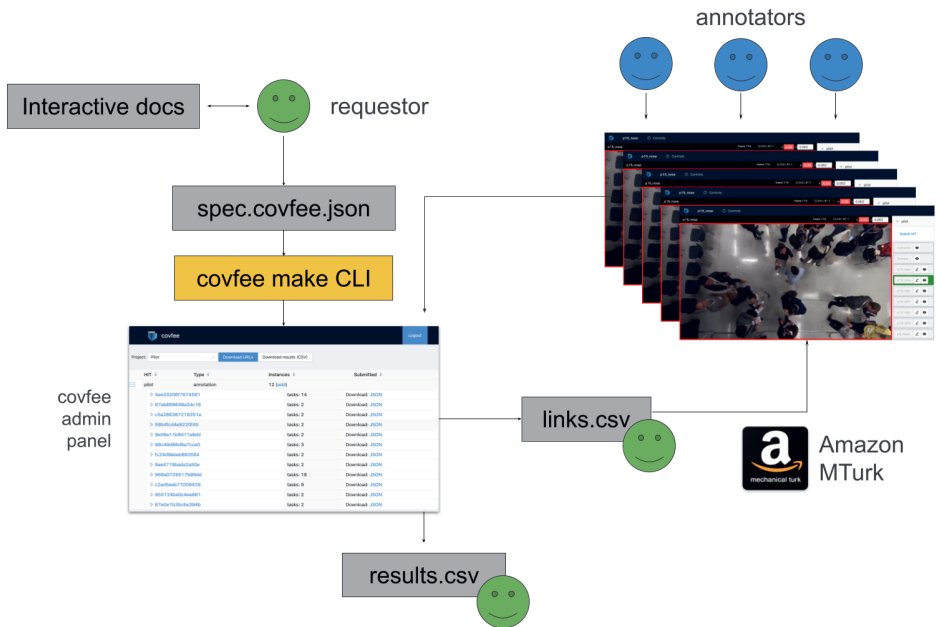


Figure 6.3: Covfee is designed to map a JSON specification into an online interface, meant to be replicated and shared online. In the basic workflow, the requester uses the Covfee documentation to create the specification, which is used to generate the online HITs for remote annotators. The HIT URLs are accessible to the requester through the administrative panel.

The workflow to be followed by the requester can be put into a sequence of steps:

1. The requester creates a Covlee specification file. Covlee's documentation was designed to help the requester create the specifications of each task.
2. The requester runs Covlee to validate the specification and generate the Covlee HITs from it. If the requester made a mistake in the specification, friendly error messages are returned indicating why and where the specification is invalid. Once a valid specification is provided, the requester can now enter Covlee's admin panel and obtain anonymized links to each HIT. A CSV file with all the links can be downloaded to be uploaded to Amazon MTurk or otherwise shared with annotators.
3. The requester may keep track of the annotation process using the admin panel. At any time it is possible to download the raw annotations in JSON and CSV formats.

For more information on the use of Covlee, please refer to Covlee's online documentation [343].

6.3.3 DATA PRIVACY AND SECURITY

Covlee deals with two kinds of potentially sensitive data: the dataset to be annotated (which could contain sensitive information about the data subjects); and the annotator responses, which could include personal information about annotators or data subjects.

Regarding the dataset, Covlee secures access to HITs via URLs containing a hash generated from a secret key. Hash URLs offer protection against scraping of the HIT links, resulting in unauthorized access to datasets while preserving the convenience of using URLs to share HITs. Using hash URLs is a standard practice for sharing documents online, with the drawback that any person with the URL may access the HIT. This is however an acceptable and somewhat necessary trade-off, given that annotators in crowd-sourcing platforms do not expect to need to create a user account on a third-party website to complete their task.

Covlee additionally provides support for data access control via required forms that must be filled in by the annotators before getting access to the data. Data access control is useful for datasets that are not publicly available on the internet, but require agreement with an End User License Agreement (EULA) on the part of the annotator. An EULA is put in place for these datasets to ensure that any person with access to the dataset agrees to the conditions stated in the agreement, which often include measures for protecting the privacy of data subjects. Many social interaction datasets are available only under an EULA [89].

Regarding sensitive responses from annotators, consent elicitation is necessary under the European General Data Protection Regulation (GDPR) when the information requested from annotators includes non-optional sensitive information. Although the GDPR is European law, it applies to the handling of data from European citizens and residents regardless of location and is considered a global reference for data protection legislation. Covlee supports consent elicitation through the same mechanics of required forms, where the user must provide their consent before proceeding with the annotation process.

6.3.4 CROWD-SOURCING SUPPORT

Covfee was created with the goal of supporting the crowd-sourcing of tasks. In contrast with a non-crowd-sourced setting, where the annotators may often be given instructions in person or via video call, in the crowd-sourcing setting communication with the annotator is generally one-way. Annotators expect to be directed to a self-contained human intelligence task (HIT) to be completed normally in a few minutes, before returning to the crowd-sourcing platform. Maximizing information flow through clear, easy-to-follow instructions and means to obtain feedback from annotators are therefore key to support this setting.

Furthermore, crowd-sourcing platforms must interface with Covfee to validate the completion of a HIT. Here, we focused on supporting integration with a) Amazon Mechanical Turk (the most popular crowd-sourcing platform) and b) Prolific, a growing platform with a focus on research studies.

Important features in Covfee that make it possible to run crowd-sourced annotation flows efficiently are:

Support for rich instruction pages A special type of task (Instruction task) can be used to provide detailed instructions in Markdown/HTML (including video tutorials). Additionally, any task in Covfee may contain tooltips to emphasize instructions or other information relevant for the annotator.

Questionnaire support Questionnaire tasks can be used to request non-continuous feedback from participants via free text boxes, buttons, sliders, and other static form elements.

Support for automatic qualification tasks For continuous tasks, a HIT may be opened only if the annotator demonstrates a certain level of ability on a shorter qualification task. A usual qualification task consists of a short sample drawn from the dataset, on which annotators are asked to follow the annotation process to be followed on the full HIT. Covfee allows the requester to easily implement a validation method, typically to compute an error between the obtained annotations and some gold standard (eg. annotations performed by the requester), allowing for some level of discrepancy (typically set empirically). Qualification tasks have been shown to improve the quality of the annotations that can be obtained in major crowd-sourcing platforms [258].

Completion codes and redirects Covfee implements integration between the Covfee platform and the crowd-sourcing platform via completion codes associated with each HIT. The completion code may be generated by Covfee and provided to the requester (following the Amazon Mechanical Turk system) or manually provided by the requester (following the Prolific system). The completion code is shown to annotators on successful completion of their HIT, to be entered by them in the crowd-sourcing platform as proof of completion. Covfee also supports redirecting annotators to external URLs on completion of their HITs.

Admin panel The admin panel, only accessible by the requester, helps keep track of progress and allows easy bulk download of HIT URLs for use in crowd-sourcing platforms.

6.3.5 EXTENSIBILITY

Covfee achieves its role as a framework, rather than simply a tool, thanks to the task-oriented class design; a user can create new Covfee tasks easily by sub-classing an existing base class. Javascript objects are available for developers to interface with. For example:

Covfee takes care of annotation recording . Covfee has methods for submitting data to the server and for reading data back from it. Data storage and client-server communication are abstracted away by Covfee. Continuous annotations are timestamped, buffered and sent to the server in chunks to minimize the risk of data loss. In addition to continuous data, tasks may submit timestamped logs of auxiliary events, like, for example, the resizing of a window or the pausing of a video. These logs may be used to collect annotations at non-regular intervals or to collect analytics with the purpose of improving the Covfee task. Non-continuous task responses may also be recorded.

Covfee's key manager makes it easy to attach event handlers to keyboard and gamepad key presses. This is especially important for continuous annotation tasks, many of which must react to button presses.

Access to Covfee's admin panel which allows to keep track of progress and download annotation results and HIT URLs easily.

Reusability Covfee tasks are modular and configurable via the JSON specification and could be incorporated as part of Covfee to be reused by others.

Covfee's socket.io module allows the implementation of multiparty tasks, where multiple subjects take part in a task at the same time. The main use case for multiparty features is not annotation but the recording of live online interactions (written, audio, or audiovisual) with the ability to query subjects at any point or request their live feedback.

6.4 CASE STUDIES

To illustrate the potential of the Covfee framework we present two case studies showcasing two custom annotation techniques: keypoint annotation and social action and confidence ratings.

6.4.1 CASE STUDY I: KEYPOINT ANNOTATION IN GROUP INTERACTION SETTINGS

In this case study, we focus on the task of labeling body joints or skeleton keypoints, particularly in the context of social interaction settings where precise, smooth annotation of keypoints over time is crucial. Manual keypoint annotations are particularly useful in the labeling of dense crowded scenes observed from the top-down view where interpersonal occlusion is minimized at the expense of more self-occlusion and more extreme perspective distortion effects. Due to the bad performance of pose estimation methods in top-down videos, automatic extraction of body keypoints is often not an option in social interaction datasets.

To implement keypoint annotation continuously, the first challenge is the difficulty of following body joints in real time, with a mouse or other signaling device. Different body parts have different motion characteristics. For example, hands and upper-body joints are used for gesturing, which can be characterized by sudden changes in velocity and acceleration, while shoulders exhibit smoother movements, and feet can be static for long periods when subjects stand still. Being able to annotate all of these accurately is vital for characterizing body movements in relation to speech. While annotating the video in slow motion would likely improve accuracy, we would like to avoid making the annotation process significantly longer. An ideal case would be if the video could be slowed down or sped up dynamically according to the speed of the keypoint that is being annotated. While this could be considered rather a chicken and egg problem since we do not yet know the speed of the object we intend to track, we propose a method below that provides a solution to this problem.

METHOD

Covfee solves the problem of continuously annotating keypoints via a new annotation technique, which involves automatically adjusting the playback rate of the video in real time, according to the magnitude of the optical flow around the mouse cursor. We thereby leverage the fact that the annotator will be pointing the cursor at the keypoint of interest and use optical flow magnitude as an approximation to the speed of the target keypoint. The video playback rate is adjusted such that it is higher when the optical flow is high and lower when the optical flow is low around the mouse cursor. This has the effect of slowing down the video when the joint being tracked moves fast and speeding up the video for slow-moving or static joints. It allows the users to annotate slow-moving joints at multiples of real time rate (eg. 4x playback rate), and fast gestures at fractions of it (eg. 0.1x playback rate) on the fly without additional user intervention.

Concretely, for a cursor position x, y (in pixels) at frame f , an $N \times N$ neighborhood in the vicinity of (x, y) is considered such that the playback rate at frame $f + 1$ is given by:

$$\hat{r}_{f+1} = C \sum_{i=x-N/2}^{x+N/2} \sum_{j=y-N/2}^{y+N/2} |O_{f,i,j}|$$

where $O_{f,x,y}$ is the optical flow vector for frame f at image location (x, y) and C is a constant. The best value of N depends on the video being annotated and is a configurable parameter.

This rate is additionally bounded to prevent extremely low or high playback rates and a user-controllable multiplier C_u is added to allow the user to control the overall playback rate:

$$r_f = C_u \max(r_{min}, \min(r_{max}, \hat{r}_f))$$

This can only be implemented efficiently in an online setting if the flow computation is done offline and only the local averaging is calculated in the user's machine. For this, Covfee makes use of a pre-computed optical flow video, which is processed in the browser making use of a Javascript version of OpenCV.js [344].

STUDY

This study presents results from applying continuous keypoint annotation, implemented in Covfee, to the annotation of keypoints in a human interaction dataset recorded during a professional social networking event. We start by comparing Covfee to a traditional, non-continuous approach using the CVAT tool on a small subset of the dataset, with annotation time and agreement as the main variables of interest. Second, we analyze the application of Covfee to the complete dataset.

The dataset used, among other modalities, contains top-down video recordings from 48 subjects, interacting freely at the same time, as shown in Figure 6.1. The interaction space was recorded by 8 cameras for 45 minutes.

Our comparison consisted of the annotation of body joints for two data subjects in the same 20s video by two sets of three annotators: one set used CVAT, and the other used Covfee. Annotators who used the continuous method were recruited from the Prolific crowd-sourcing platform, without any filtering, and provided with a link to a HIT in Covfee. Because the CVAT tool does not implement support for crowd-sourcing, annotations in the CVAT condition were performed locally by three of the authors. No annotators had previous experience with any of the tools and work conditions were not controlled. Although crowd-sourced workers may have had previous experience in other kinds of annotation, we think the difference between our continuous keypoint annotation task and most crowd-sourced tasks is significant enough to make this experience unlikely to be a source of bias.

All annotators were provided written instructions to label the left shoulder, right shoulder, the center of the head, and a point in the direction of the gaze of the data subject (ie. in the direction of the nose). The goal with this last point was not to measure its precise location in pixels but to use it to obtain a head orientation vector. Local annotators were asked to measure their total annotation time for CVAT. For Covfee, the time was acquired from the difference between the timestamps that Covfee adds to each data point.

In the case of CVAT, frames were annotated every second and linearly interpolated in between. For Covfee, the method in Section 6.4.1 was used with parameters $N = 20$ (pixels), $r_{min} = 0.1$, $r_{max} = 4$. The video was pre-processed by denoising with the *hqdn3d* filter in FFmpeg [319] with a temporal luma strength $luma_tmp = 30$.

The annotators reported lower annotation times on average for the continuous approach (7.4 minutes) compared to taking between 17 and 25 minutes for the CVAT annotations. We compared the annotations for head and shoulder key points by computing the average Euclidean distance in pixels between time-corresponding annotations. We averaged this discrepancy for all pairs of annotators. On average, our continuous annotation approach resulted in lower discrepancy (18.7 ± 10.0) when compared to the use of CVAT (22.9 ± 12.7), although within standard deviation.

The same was true when we measured the discrepancy in the orientation of body and head in degrees (7.9 ± 4.8 for Covfee vs 9.9 ± 10.7 for CVAT). Body orientation was computed by taking the vector between both shoulder points and head orientation was computed by taking the vector between the head keypoint and the gaze direction keypoint. Table 6.1 shows the errors measured per keypoint and annotation times in Covfee and CVAT. Annotation times for CVAT were not measured per keypoint. However, given that the CVAT annotations were image-based with a fixed interval between images, we expect annotation times to be roughly equal across keypoints (5.25min on average). It is

Body joint	CVAT disc.	Covfee disc.	Covfee time	CVAT time
Head (px)	12.6 (7.9)	14.4 (12.0)	4.2min	5.25min
Left shoulder (px)	21.4 (11.1)	19.7 (6.9)	1.5min	5.25min
Right shoulder (px)	34.5 (19.1)	22.1 (11.2)	1.7min	5.25min
Head orientation (deg)	11.4 (12.8)	7.3 (4.0)	N/A	N/A
Body orientation (deg)	8.3 (19.9)	8.4 (5.6)	N/A	N/A

Table 6.1: Results of the comparison between Covfee continuous annotation and CVAT in the annotation of body joints. Values in parenthesis are standard deviations. Annotation discrepancies are averaged distances between corresponding annotations, over all pairs of annotators. Lower discrepancies indicate higher agreement. CVAT times are averaged since only totals (for all four annotated keypoints) were reported and we expect annotation times to be roughly constant for the four body joints. Note that the head and body orientation are derived values, hence no time is reported.

particularly noteworthy that the head keypoint took on average significantly longer to annotate using Covfee, which is likely to be due to the head moving more rapidly during these segments. Even though our annotator sample was too small to measure differences in discrepancy across conditions, we are confident that the large (significant) differences in annotation time generalize to other situations. Even if true annotation quality were to be lower for the continuous case, we think the gains in annotation time are enough to make this an attractive approach for large-scale annotation.

Given the results of the previous comparison, we proceeded to use Covfee to annotate body joints in the complete dataset. A total of 17 body keypoints (joints) were annotated for each subject, for a subset of 16 minutes of the dataset. This was equivalent to more than 218 hours of single-keypoint tracks.

Having an *occlusion* signal for each keypoint is important in the training of pose estimation methods (see Section 6.2.1). To support this important signal, we integrated a binary occlusion label into our technique by recording an additional key press. We asked annotators to hold down a keyboard key (while following the keypoint with their cursor) when the target joint was occluded. If the joint was still within the frame despite being occluded, annotators were asked to follow it approximately by inferring its location. While these annotations would in principle be *filtered out* of the training process, asking annotators to infer location in this way enables them to maintain the continuity of the annotation. Additionally, though not standard practice, pose estimation methods could be trained to estimate occluded keypoints in addition to visible ones.

Adding this additional input made the annotation process slightly more involved, although in our pilot tests we did not notice any cognitive load issues with simultaneously following a keypoint with the mouse and annotating occlusion with the keyboard. Figure 6.4 shows the mean occlusion levels annotated over the image plane, averaged over our multiple videos. These plots use the same color scale and show the spatial variation in occlusion levels for body keypoints: head and feet. Continuous occlusion annotations allowed us to obtain a richer description of the skeletal data without increasing annotation time.

In summary, this case study of keypoint annotations showcases how Covfee supports a continuous annotation procedure that provides richer and better quality information

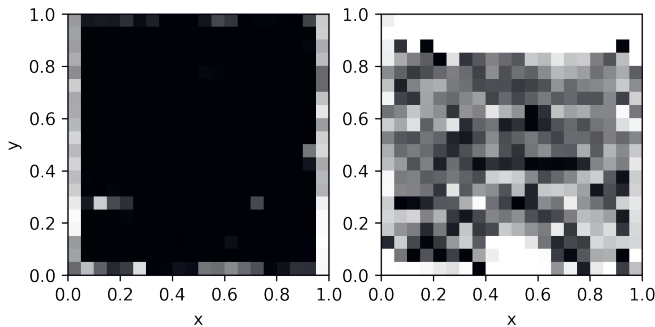


Figure 6.4: Plot showing the distribution of our occlusion annotations for head (left) and feet (right) keypoints. White indicates high occlusion and black low occlusion values. The head keypoint, being visible from most angles shows little occlusion while the feet tend to be more occluded when near the edges of the frame and show overall higher occlusion values.

about human body movements during socializing. This is in part due to the time-efficient nature of the continuous annotation process, which allows for additional annotations to be made that can help us to understand and characterize better the relationship between the phenomena that are being labeled and the annotation noise.

6.4.2 CASE STUDY II: SOCIAL ACTION ANNOTATION

The annotation of speaking status is particularly key in automated social interaction inference tasks. However, recording audio of people in real life settings can be very privacy invasive. Fortunately, from past efforts [89] we know that it is possible to annotate speaking status from video only with some degree of annotator agreement sufficient for training machine perception systems [316], although short back channels can be difficult to capture [89]. Acceptable inter-annotator agreement from video only can be explained by the fact that when humans speak, their vocal behavior is often accompanied by linguistically related body movements such as gestures[207].

This Section describes a case study about the annotation speaking status from video in a large social interaction dataset. The action of speaking was annotated using binary continuous annotation, where annotators were asked to hold down a keyboard key whenever they perceived speaking to be happening in the video.

In real life in-the-wild settings, videos may not always capture the subject of interest very clearly. The person may be partially occluded by others in the scene, they may have their back to the camera, or their face may not be visible. Access to multiple viewpoints of the data subjects is desirable to offset these challenges. This is however not a complete solution as in some cases none of the views may offer a suitable view of the subject of their speaking behavior could be hard to discriminate. This is a common situation with data recorded in real-life settings when intrusive sensing is avoided to preserve the naturalness of the interactions. To capture this uncertainty it would be of great benefit to know the confidence of the annotator in their judgement. To this end, we obtained continuous confidence ratings by asking annotators to indicate the degree of confidence that their action assessment (either speaking or non-speaking) was correct. In a training stage, such

a confidence signal can be used to give less weight to data samples or segments for which the annotator had low confidence on being correct.

METHOD

Covfee supports action annotation via an interface for binary continuous annotation. The annotator can control the binary status of the annotation via a keyboard key: *true* if the key is pressed; *false* if it is not. Visual feedback is provided when the key is pressed.

Confidence annotations were also performed continuously in Covfee using an interface designed in general for continuous-value annotations. In this interface, the users can control a vertical slider using their mouse. The vertical position of the slider follows the cursor's vertical position. The continuous value of the slider indicator (in the range $[0, 1]$) was recorded in Covfee.

STUDY

Our study on actions is based on the data obtained from the annotation of a large dataset (see Section 6.4.1) for speaking. Annotators were part of a larger group who worked on the annotation of our dataset, both for keypoints and speaking status. We selected conscientious annotators for this group via a short qualification task consisting of keypoint annotation only, and revised manually via playback of their annotations, but otherwise no special selection of annotators was done, nor did we control their working conditions. Annotators from the larger group worked on action annotation based on their availability when this phase of the project was reached.

In the action annotation stage, annotators were instructed to annotate the speaking status of all subjects in the scene, and to continuously *annotate* their confidence in their judgment about speaking status, per the method described above. To offset the issue of lack of visibility of the target subject, we gave annotators access to several side-elevated views of the subjects, from which they could pick the best one.

Computing turn lengths from the obtained speaking status annotations revealed that a high proportion of turn lengths were below one second in length, suggesting that we were able to capture quick turns, and potentially back-channels. Although we do not have access to speaking ground truth to verify it, our confidence annotations give us annotator ratings of the degree of certainty in their inferences. Figure 6.5 plots the turn lengths obtained from our annotations against the average confidence annotated (by the same annotator) during the corresponding turn. The plot does not reveal a clear trend, suggesting that confidence was not heavily dependent on turn length. It is likely that other factors like visibility may influence annotator confidence more.

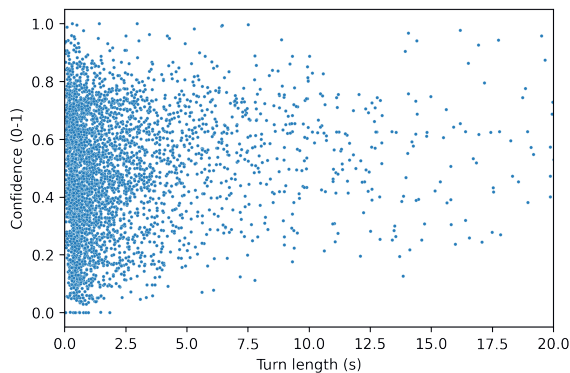


Figure 6.5: Plot showing the correlations between annotated turn lengths and mean annotation confidence during the turn for our speaking status annotations.

6.5 CONCLUSION AND DISCUSSION

In this paper, we have presented Covfee, a new web-based framework with the goal of supporting the study and use of continuous annotation in human behavior data. Although continuous-time annotation has long been used for affective dimensions, we present Covfee as a general framework, capable of supporting both these established techniques and new continuous annotation techniques. The motivation to support novel continuous annotation techniques for human behavior datasets (eg. for body joint and action annotation) comes from the potential to improve the time-efficiency of the annotation process when single subjects are annotated for long periods of time (minutes to hours); and the suitability of continuous techniques for annotations that rely heavily on temporal context.

We have laid out the design decisions and main features of our framework, aimed both at basic users without knowledge of web development who wish to use existing tasks out of the box, or those with web development skills who wish to build new annotation techniques on top of Covfee. We started by explaining the workflow for requesters to use Covfee, which revolves around a specification file describing the HITs to be created. Covfee processes the specification file to create the annotation interfaces specified in it, and makes HITs available under a secure URL. We go on to explain how the tool supports data privacy and security for the annotation of potentially personal or sensitive data. We emphasize the design choices and features that make Covfee suitable for a crowd-sourcing setting and lay out the features that make Covfee a framework, applicable to the implementation of new continuous tasks.

We presented two case studies applying continuous annotation (and Covfee) to keypoint and action annotations in a social interaction dataset. Our study on keypoint annotation showed an improvement in annotation time without a significant difference in annotation quality. Furthermore, our continuous technique allowed us to annotate keypoint occlusion in the same pass. This auxiliary signal is very relevant for the pose estimation task since methods are usually fed only the set of visible keypoints and the occlusion signal is used to filter the input to the method. In traditional techniques, occlusion annotations are limited in time resolution by the frequency at which frames are annotated and cannot be

interpolated between annotated keyframes like continuous values can. Our continuous technique, in contrast, provides a higher-resolution signal indicating when each keypoint becomes visible or occluded.

Similarly, in our action annotation study, we obtained continuous-valued annotation confidence labels together with our binary speaking status annotations, although this time in a second pass over the data. Confidence signals give the researcher access to a measure of uncertainty in the data labeling at each moment in time, without having to label the data multiple times to obtain an agreement-based measure. One important question for future work is how well ratings of confidence from a single annotator approximate agreement measured from multiple annotators. Continuous-value annotations are also known to be affected by bias when interpreted absolutely (see Section 6.2.3). The extent to which this bias affects ratings of confidence including ours, and the best way to elicit and interpret confidence annotations are also open questions.

Although our study on actions did not involve an annotation time comparison with traditional action annotation techniques, we think that continuous annotation may also be a more time-efficient way to annotate most human actions since it can be done in real time, or even fast motion without the need to pause for labeling. Importantly, we think that time efficiency should not be the only consideration when deciding for or against a continuous technique. In our experience, the suitability of continuous annotation for actions depends on the desired precision, frequency, and context-dependency of the actions being annotated.

Regarding temporal precision, this is usually a function of the research questions being investigated. In human behavior research, certain research questions involve the precise localization of action onsets and offsets, where onset and offset are reasonably well-defined and observable. Studies on the internal structure of gestures and laughter episodes, for example, make use of fine-grained temporal segmentation [345]. In this case, continuous annotation alone might not be a suitable solution given the annotation delays involved. Continuous annotation may however still be useful when the annotation task can be separated into two steps; first, a continuous localization step (where actions are localized roughly in time) followed by a second precise temporal segmentation step, where a precise coding scheme is applied. In other words, at present, we do not envision continuous action annotation as a complete solution for behavioral coding, but rather as a method for rough time-localization of phenomena of interest. In many machine learning applications, however, precise localization of action boundaries and action segmentation is not a requirement and robust machine learning methods or correction techniques have been proposed to mitigate the effects of delay in continuous annotations [117, 119].

Regarding action frequency, continuous annotation provides greater time improvements the more frequent the target actions are. For extremely sparsely-occurring actions the time gained from continuous annotation becomes lower, as even in the non-continuous case, annotators would spend most of the time watching the media, and less time annotating. However, many actions of interest in human behavior research are frequent enough to benefit greatly from continuous annotation in terms of time efficiency. In social signal processing and affective computing, actions such as speaking, gesturing, laughing, and other common actions in a social context are often annotation targets.

Finally, concerning the temporal-context-dependency of the actions, we think conti-

nuous approaches are advantageous for most actions occurring in a social context because they enable the annotator to follow the flow of what is occurring in the interaction without interruption. Annotation of actions or situations such as *use of humor* or *enjoyment* requires a complex context-based judgment on the part of the annotators. Such context-heavy constructs are however common annotation targets in communities working with in-the-wild data such as social signal processing or affective computing.

Given these trade-offs we argue that continuous annotation is much more useful for action annotation than its current usage would suggest.

It is important to highlight once again, however, that continuous annotation may not be suitable for every problem. The standard technique of bounding box annotation, for example, does not straightforwardly translate to the continuous case since it is not clear how an annotator would control the location and dimensions of the bounding box continuously. This task is also hard to decompose into single-point annotation tasks since the corners of the box may not correspond to any meaningful keypoints in the scene. We cannot rule out, however, that new creative techniques will make it possible to perform such annotations continuously. Hybrid techniques where manual annotation is aided by models are not new and the application of such approaches to continuous annotation may open the door to new breakthroughs in annotation efficiency.

In general, Covfee has the long-term goal of dramatically improving the time and effort necessary to collect and annotate human behavior data online. It was born out of the need for a web annotation platform flexible enough to accommodate the high diversity and specificity of annotation needs present today. We expect that all of the design decisions made to support this goal will enable the adoption of Covfee as a platform for a) the implementation of existing annotation techniques such as those traditionally used within the affective computing community, b) experimentation with novel annotation techniques for vision tasks, such as the two techniques presented in this paper and c) developments in other fields such as the annotation of audio or other time series.

Appendices

Covfee: an extensible web framework for continuous-time annotation of human behavior

6.A COVLEE SPECIFICATION EXAMPLES

Here we present an example of a Covfee specification file for reference purposes. Using Covfee as a requester involves writing such a specification file:

```

1 {
2   "id": "1d_annot",
3   "name": "Continuous annot sample",
4   "email": "example@example.com",
5   "hits": [
6     {
7       "id": "1d_annot",
8       "name": "1D annotation sample",
9       "repeat": 2,
10      "tasks": [
11        {
12          "type": "ContinuousKeypointTask",
13          "name": "Head",
14          "media": {
15            "type": "video",
16            "url": "$$www$/myvideo.mp4",
17            "resolution": [1920, 1080],
18            "fps": 25
19          }
20        }
21      ]
22    }
23  ]
24 }
```

Note: the exact schema of the specification file may vary with new Covfee releases.

Figure 6.6 shows the output of validation of the Covfee specification above, with a mistake in the name of the task (*ContinuousKeypoint* instead of *ContinuousKeypointTask*) to showcase the kind of error messages that Covfee returns to the user.

```

✓ 1 covfee project files found.
✓ Read covfee file annotation.covfee.json.
✗ Error validating project "Continuous annot sample" in annotation.covfee.json.
```

Error in project["hits"][0]["tasks"][0] for object:

```

{
  "type": "ContinuousKeypoint",
  "name": "Head",
  "media": {
    "type": "video",
    "url": "$$www$/sign.mp4",
    "resolution": [
      1920,
      1080
    ],
    "fps": 25
  }
}
```

Error: Invalid value 'ContinuousKeypoint' for property 'type'. Please make sure you are using a supported value.

Figure 6.6: A validation error generated by Covfee indicates the location of the error within the specification and the reason for it being invalid.

6.B COVLEE WEB APPLICATION

HIT	Instances	Submitted	Links	Edit
+ G0_HIT1	7		Download	Edit
- G0_HIT2	5		Download	Edit
> b70cc09f958f578		tasks: 1	Download .JSON CSV	
> 24c6afa02436fd25		tasks: 1	Download .JSON CSV	
> ec3023084dc965b1		tasks: 196	Download .JSON CSV	
> 99c6253ff680a325		tasks: 196	Download .JSON CSV	
> 1f2ad3a34e75020		tasks: 1	Download .JSON CSV	
- G0_HIT3	5		Download	Edit
> 94096acaf55970a		tasks: 1	Download .JSON CSV	
> eda1bf4c0b1962f1		tasks: 196	Download .JSON CSV	
> 97136a9c51b81bab		tasks: 196	Download .JSON CSV	
> a286a98e7a33baac		tasks: 196	Download .JSON CSV	
> 2d07f6bc4c29b478		tasks: 196	Download .JSON CSV	
+ G1_HIT1	5		Download	Edit
+ G1_HIT2	6		Download	Edit
+ G1_HIT3	5		Download	Edit
+ G2_HIT1	3		Download	Edit
+ G2_HIT2	5		Download	Edit
+ G2_HIT3	3		Download	Edit
+ G3_HIT1	3		Download	Edit
+ G3_HIT2	3		Download	Edit
+ G3_HIT3	4		Download	Edit
+ G4_HIT1	4		Download	Edit

Figure 6.7: The admin panel allows for tracking of HITs, downloading results, and accessing URLs for dissemination.

Consent

automatically. Withdrawal after completion of the experiment is not possible as the data will be stored completely anonymously (i.e. we cannot link it back to your identity).

5. Funding for this project is provided by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.506.

6. For issues, concerns or complaints please contact: The responsible researcher: Jose Vargas Quirós [j.vargasquiras@tudelft.nl +31626520033]

The institution of the researcher: Delft University of Technology info@tudelft.nl +3115 2789111

The funding source of this project: NWO The Hague nwo@nwo.nl +31 (0)70 3440040

I have read and understood the study information dated [DD/MM/YYYY] or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

Yes

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

Yes

I understand that taking part in the study involves providing data settings or annotations which will be recorded but not linked to my identity, along with data about my interaction with the web interface.

Yes

I understand that information I provide will be used for scientific publications.

Yes

I understand that personal information collected about me that can identify me, such as [e.g. my name or where I live], will not be asked nor stored.

Yes

I give permission for the annotations that I provide to be archived in networked storage at Delft University of Technology in anonymized form so it can be used for future research. I understand that data will not be made public but may be shared with researchers under an End User License Agreement (EULA).

Yes

[Next](#)

Figure 6.8: Consent forms can be specified directly in Covfee. They can act as a requirements before annotators get access to the data.

The screenshot displays the 'General Instructions' section of a Covfee HIT. On the left is a navigation menu with a 'Consent' dropdown and a list of sections including 'General Instructions', 'Reaction Time', and several 'Example' tasks for 'Recognition' and 'Rating'. The main content area features several instructional boxes: a red 'Please use Chrome' warning, a 'General instructions' section with a welcome message and task details, a green 'Important' box with three key instructions, an 'Example tasks' section with a blue 'Next' button, and a 'Next' button at the bottom.

Please use Chrome
Some of the tasks in these experiments only work on Google Chrome. Please do not proceed on another browser.

General instructions
Welcome!
The following sections consists in rating short recordings (approx 7 seconds each) of laughter in social interaction. Although many of the recordings contain laughter, not all of them do, and your job will be to indicate which ones you think contain laughter, and which ones do not. **For each short video** there are two tasks to complete:

1. Recognition: you will have to press a key (q) when you perceive laughter to be happening in the video.
2. Rating: if you thought there was laughter, we ask you to rate the intensity of the laughter and your confidence that there was or was not laughter.

Important

- We want to capture your gut feeling about laughter. Do not worry about your work being perfect. Many of the videos are challenging.
- Make sure to maximize your browser window before you start annotating.
- Make sure you have headphones at hand. You will need to listen to audio.

The whole experiment is divided into four sections:

- **Examples (3 videos)**: These are three videos with clear laughter for you to familiarize yourself with the task.
- **Video-only (40 videos)**, where the videos contain **no audio**. You'll have to pay attention to the body movements to try to determine if there is laughter.
- **Audiovisual (40 videos)**, where the videos contain audio.
- **Audio-only (40 videos)**, where you can only listen to the laughter but not see the person.

You will also encounter a few tasks designed to measure your reaction time.

Example tasks
We prepared three easy examples for you to learn the task. The results will be discarded. Please use them to practice and make sure you understand the tasks.
Click Next to start the example task!

Next

Figure 6.9: Rich instructions in Markdown / HTML format can be specified directly into a Covfee HIT. Providing good instructions is key for novel annotation techniques that annotators are unfamiliar with.

7

IMPACT OF ANNOTATION MODALITY ON LABEL QUALITY AND MODEL PERFORMANCE IN THE AUTOMATIC ASSESSMENT OF LAUGHTER IN THE WILD

7

Although laughter is known to be a multimodal signal, it is primarily annotated from audio. It is unclear how laughter labels may differ when annotated from modalities like video, which capture body movements and are relevant in in-the-wild studies. In this work, we ask whether annotations of laughter are congruent across modalities, and compare the effect that labeling modality has on machine learning model performance. We compare annotations and models for laughter detection, intensity estimation, and segmentation, using a challenging in-the-wild conversational dataset with a variety of camera angles, noise conditions, and voices. Our study with 48 annotators revealed evidence for incongruity in the perception of laughter and its intensity between modalities, mainly due to lower recall in the video condition. Our machine learning experiments compared the performance of modern uni-modal and multi-modal models for different combinations of input modalities, training, and testing label modalities. In addition to the same input modalities rated by annotators (audio and video), we trained models with body acceleration inputs, robust to cross-contamination, occlusion, and perspective differences. Our results show that the performance of models with body movement inputs does not suffer when trained with video-acquired labels, despite their lower inter-rater agreement.

7.1 INTRODUCTION

Laughter is traditionally associated with its characteristic vocalization (ie. the sound of laughter). In research too, its vocal manifestation has received the most emphasis.

Nonetheless, laughter is a multimodal phenomenon. Darwin presented a curious depiction of excessive laughter: “The whole body is often thrown backward and shakes, or is almost convulsed; the respiration is much disturbed; the head and face become gorged with blood; with the veins distended; and the orbicular muscles are spasmodically contracted in order to protect the eyes. Tears are freely shed.” [346, p.208]. This depiction makes reference to multiple characteristic manifestations of laughter: the facial movements of laughter, the full-body movements of laughter, and the physiological changes of laughter.

Following this premise works in social signal processing [96, 347] have delved into the problem of automatically detecting and classifying laughter from audio, video, and audiovisual recordings of its manifestations. Annotation is a key step in these studies. The first step in annotation of naturally occurring laughter usually involves the temporal localization, or segmentation of laughter (from its context). Next, laughter segments or episodes are categorized or otherwise rated. Functional or formal categorizations are the most common, but no consensus coding schemes exist for either of these tasks. Laughter intensity is also a common variable of interest that has been rated in multiple studies [5, 34, 106, 348–351]. Mazzocconi et al. [10] have linked laughter intensity directly to the meaning of laughter, as an indication of the magnitude of a positive shift in arousal caused by the laughable (the object of laughter) in the laughing subject.

Nevertheless, the emphasis on the vocal manifestations of laughter translates strongly to its annotation, where laughter has most commonly been annotated from audio or audiovisual face recordings, by third-party observers [352–355]. Less commonly, laughter has been annotated from body movements alone, using video. This has been done in in-the-wild datasets of *mingling crowds* recorded in real-life events [89], such as the dataset in Figure 7.1. In these datasets, audio recordings are commonly not available, due to the technical and logistic difficulty, and privacy challenges when equipping each study participant with a microphone. In-lab studies of the body movements of laughter have also often opted for video-only annotation of laughter, to align with the target task under study.

However, it is unknown if video labeling of laughter has a relevant effect on annotation quality, and how annotations acquired in this way differ from the more common audio-based and audiovisual annotations. The same is true for audio-based labeling: the benefits of including video at annotation time have not been verified. In other words, the consequences of the choice of annotation modality have received little attention in research. Furthermore, it is unclear whether ratings of the intensity of laughter can be expected to be congruent across modalities, a question with direct implications in the interpretation of laughter [10].

While inter-rater agreement is an important dimension of annotation quality, higher annotation agreement does not necessarily imply superior model performance. The question of how annotation modality impacts model performance is, therefore, a separate, yet also unexplored question.

Answering these questions is important both for the interpretation of previous work focusing on a single modality and for informing annotation choices in future work. In this work, we take a first step in that direction by studying laughter annotation across modality conditions. First, we investigate how the human ability to detect, segment, and estimate the intensity of laughter (three foundational tasks in laughter work) differs with and without access to video and audio. Due to the difficulty of collecting audio in in-the-wild mingling settings [89], we use an in-the-wild mingling dataset containing full-body



Figure 7.1: Screenshots from the four elevated views in our dataset of free-standing interactions used in this work.

motion information. Data was collected during a real-life event and contains naturally occurring laughs (Figure 7.1). Body movements of laughter (eg. shaking, swaying, arm and feet movements) can be observed in the videos, but access to facial features is limited due to occlusion. These factors, along with the diversity of camera angles, and distances to the camera make the in-the-wild mingling setting one of the most challenging scenarios for laughter perception, especially from video.

Second, we study how labels acquired under different modality conditions affect the performance of machine learning models for laughter detection, segmentation, and intensity estimation. We pay special attention to the question of whether video-acquired annotations result in performance comparable to that of audio and audiovisual annotations. Naturally, the input modality of the model itself plays an important role here. We compare models trained on the same input modalities used to annotate: video, audio, and audiovisual. Additionally, we included accelerometer readings from chest-worn wearable devices (worn by many subjects in our dataset) as an additional model input. Such accelerometer readings have been used in previous work for the detection of multiple social actions such as speaking [81, 95, 128], with performance competitive and often superior to that of video. Furthermore, wearable accelerometers have privacy and scalability advantages due to their low cost and their ability to capture information from the device wearer alone. We hypothesized that acceleration would have a behavior similar to video since both modalities capture primarily body movement information. However, we expected acceleration to better capture laughter intensity when compared to video due to its orientation invariance, and to it not being affected by occlusion and cross-contamination like video is. Our contributions are the following:

- We present the first human study of laughter annotations across annotation modali-

ties, comparing three conditions of interest in previous work: audio-only, video-only, and audiovisual. We studied the three annotation tasks of laughter detection, time-localization, and intensity rating.

- We present a cross-modal analysis of annotations via inter-annotator agreement within and between annotation conditions: video-only, audio-only, and audiovisual. We obtained insights important both for the interpretation of previous work annotating on a single modality and for informing annotation choices in future work.
- We investigated the effect of annotation modality on machine learning model performance. Mirroring the human study, we used state-of-the-art models for detection, intensity estimation, and time-localization. We implemented, trained, and evaluated models for different combinations of input modalities (audio, video, acceleration), training, and testing label modalities (video, audio, and audiovisual). It is shown that despite the lower inter-annotator agreement of video-based labeling, they may be entirely appropriate to train models for laughter detection from body movements.

7.2 BACKGROUND AND RELATED WORK

In this section, we discuss laughter annotation in research, especially in computational work towards understanding laughter. In Section 7.2.1 we start by briefly summarizing the research landscape surrounding laughter. In section 7.2.2 we discuss automatic laughter detection and related machine learning tasks. In section 7.2.3 we discuss work on laughter annotation and how laughter has been annotated in previous studies.

7

7.2.1 THE STUDY OF LAUGHTER IN INTERACTION

Laughter has been approached from the perspective of multiple scientific disciplines. Psychology, is concerned with, among others, the semantics and functions of laughter in interaction [11, 356, 357]. In biology, the evolutionary origins [15] and physiological effects of laughter [19] are the subject of study. Meanwhile, social signal processing, speech and human-agent interaction fields are concerned with automatic tasks such as laughter detection [358, 359], classification [107, 108] and synthesis [360], with datasets being created for the study of laughter in specific [325, 353, 361].

Laughter is most often analyzed as a meaningful signal in social interactions, as it is an overwhelmingly social phenomenon found to be about 30 times more likely in social situations than when by oneself [362]. To this end, drawing a parallel with the study of speech, Mazzocconi et al. [10] distinguished two broad levels for the study of laughter: 1) laughter form and context and 2) laughter's (social) meaning and function. Laughter form includes the physiology and body movements of laughter and its acoustic features; and laughter context includes its positioning with respect to speech, to others' laughter, and to its object (the laughable). Most of the work on the form of laughter is concerned with its phonetics and acoustic structure, with different coding schemes for segmentation of laughter into its constituent (acoustic) components often being used [345]. Laughter intensity has also received attention as a dimension of laughter form [5, 22, 34, 106, 348–351, 363]. Most laughter in conversations has been observed to occur at relatively low intensity [22].

Laughter form and context influence its second level of analysis: the meaning and function of laughter. Mazzocconi et al. propose the following as the meaning of laughter: “The laughable l having property P triggers a positive shift of arousal of value d within A ’s emotional state” [10, p.4], where A is the producer of the laugh. This interpretation provides a link between laughter intensity and laughter meaning. Despite the importance of laughter intensity in previous work, it is not known to what extent intensity ratings are congruent (or not) across modalities.

Laughter has been found to serve a multitude of functions at the coordination level (taking the role of punctuation [362, 364] and as a cue for topic termination [365, 366]) and at the social level to foster relationships, cooperation, and group cohesion [8].

7.2.2 AUTOMATIC LAUGHTER DETECTION, CLASSIFICATION, AND INTENSITY ESTIMATION IN THE WILD

Most research in laughter detection and classification has made use of meeting datasets and focused on the audio and audiovisual modalities. Truong et al. [354, 367] used spectral features, pitch, energy, and voicing to discriminate laughter from speech. In a series of papers, Petridis et al. investigated audiovisual laughter detection and discrimination [368, 369] from upper body meeting videos, using static and dynamic features fed into a single-layer perceptron.

There have been fewer attempts to automatically assess laughter exclusively from the video modality. Mancini et al. [348] proposed a method to estimate laughter intensity from the movement of shoulder and head keypoints in a video. More recent action recognition methods based on 3D convolutional neural networks (CNNs) [370] have not been applied and analyzed in this task.

Full body poses and acceleration have also been inputs of interest. [84, 108] showed that traditional classifiers are capable of recognizing and classifying elicited laughter from pose sequences alone.

The related task of voice activity detection (VAD) has received more attention in in-the-wild settings, with models having been proposed for the detection of speech from video alone [80, 227]. Here, a deep 3D-CNN-based model has been shown to improve over previous approaches [127]. Additionally, work with accelerometer inputs has shown that this modality holds sufficient discriminative power to improve over larger video-based methods [127, 128].

7.2.3 LAUGHTER PERCEPTION AND ANNOTATION

At its lowest level, laughter annotation is concerned with the recognition and segmentation of the form of laughter. Most of the work on the form of laughter is concerned with its phonetics and acoustic structure. Laughter is typically classified in voiced, unvoiced, and speech laughter [371] depending on the degree of engagement of the vocal chords [372]. Regarding its temporal extent, there is not a widely accepted definition of what constitutes a laughter episode. Most studies of laughter delving into its structure have relied on audio waveforms for the segmentation of laughter, typically into laughter syllables or vowels (ha) at the lowest level, followed by bouts (sequences of syllables), which are separated by inhalations [373]. Truong et al. propose a multi-level segmentation scheme to describe the structure of laughter [345]. This scheme, however, relies on audio alone.

Body movements, especially those occurring below the face, have been largely disregarded in the study of laughter form. There are, however, notable exceptions. In a perception study, Griffin [108] showed that humans are capable of recognizing laughter and even of classifying it functionally based on stick figures. The use of stick figures provided a way to isolate the body movement component of laughter. Note however, that in contrast to our work, this study was not concerned with annotation (where the goal is to use the most reliable information available) and did not analyze agreement across modalities.

In the work most similar to ours, but focused exclusively on facial movements, authors created visual, audio, and audiovisual laughter stimuli/examples from face recordings [374]. The audio contained different levels of artificial noise to make laughter more difficult to detect. 20 annotators indicated if they perceived laughter or not in these examples. The goal was to study how much the face contributes to the perception of laughter. The study reported that “visual laughter consistently made auditory laughter more audible” (ie. audiovisual laughter was easier to detect than audio-only laughter), a phenomenon also observed previously for speech perception [374]. Although this is, to the best of our knowledge, the only work to perform a cross-modal analysis of laughter perception, its findings do not necessarily generalize to our setting, where the video modality contains overall body movement information, but facial movements are not consistently available. Furthermore, being a perception study, they considered expert annotations to be ground truth but provided no analysis of inter-rater agreement.

Most studies of automatic laughter detection (see Section 7.2.2) rely on laughter annotations made from audio [359, 368] (possibly automatic like the ones in AMI [375] and SEMAINE [32]) or audiovisually [83, 84, 376]. However, studies concerned with the body movements of laughter often obtained ground truth annotations from the video modality alone. [348] rated laughter intensity from body movements alone. Cu et al. [103] annotated five affective categories of laughter from body movement, without sound. These studies, however, do not offer a comparison with audio-based annotation, and it is therefore uncertain to what extent annotations would be congruent across modalities.

7.3 OUR APPROACH

Answering our research questions requires the annotation of a large set of laughter segments with associated audio and video signals. Measuring inter-annotator agreement across conditions additionally requires that the same segments are annotated by multiple annotators. Annotations must also be done by a representative sample of annotators, large enough to ensure that individual biases do not drive the results. The first step in a study of laughter in the wild is to localize laughter in the target dataset. Ideally, a large number of annotators would each watch our complete dataset (with more than 50h of individual behavior) to find and annotate laughter episodes. This, however, would involve thousands of hours of human labor. Due to the relative scarcity of laughter in conversation in the wild, most of this time would be spent listening to speech with only sporadic laughter. Therefore, the first simplification that we adopted was to pre-localize *laughter candidates*. *Laughter candidates* are segments where the author of the study (who did the pre-annotation) perceived laughter to occur. The pre-annotation was done inclusively, meaning that in case of doubt, laughter was always annotated. These positive candidates were complemented with negative examples, where laughter was not perceived to occur, to obtain a dataset

of laughter / non-laughter candidates. The resulting dataset was used both for human annotation and machine learning experiments.

Figure 7.2 shows an overview of our study. In Section 7.4 we present the audiovisual dataset chosen. In Section 7.5 we delve into our methodology for the design of our human annotation study (Section 7.5.2); analysis of annotator agreement (Section 7.5.3); and analysis of machine learning model performance (Section 7.5.4) for classification, segmentation and laughter intensity estimation.

7.4 DATASET

Our dataset was collected during a business networking event in Delft, The Netherlands. Subjects in the experiment were members of a group organizing regular events. During the event, most of the interaction consisted of free-standing conversation (Figure 7.1). Participants were free to move around as they pleased. While the event also included several pre-planned activities including a social game and music performance, we excluded these moments and made use only of the segment of the data containing free interaction. The following data was collected during the interaction:

Body Acceleration. A wearable accelerometer sensor that was hung around the neck like a badge measured upper torso acceleration.

Individual Audio Lavalier-type microphones attached to the faces of participants recorded sound at 44.1kHz. Microphones were connected to a Sennheiser SK2000 transmitter worn around the waist area. Our audio equipment consisted of 32 microphones. These individual audio recordings were used to obtain Voice Activity Detection (VAD) labels at 100 Hz for each participant, making use of rVAD [240], a state-of-the-art unsupervised VAD method specially designed for noisy audio. 100 Hz is the fixed output frequency of rVAD and enough to capture even single syllables in languages like English [377].

Video 12 overhead cameras and four side-elevated cameras were placed above and in the corners of a video zone. Participants were informed about this video zone and asked to stay outside if they did not wish to be recorded. In this work, we only make use of the side elevated cameras, due to it being a viewpoint more familiar to observers and able to capture the face. Figure 7.1 shows a capture of the four elevated camera views.

In coordination with event organizers, it was decided that each participant would be free to choose which sensors to wear (microphone, accelerometer, or both). Of about 100 attendees to the event, 43 wore a sensor during the event. Of them, 20 were male and 23 female. The rest decided not to take part in the data collection, or could not be given a sensor due to our supply limit.

While similar *mingling* datasets have been published in the past [88, 89], our dataset was the first to contain high-quality individual audio recordings, opening the door for cross-modality studies such as this one.

7.5 METHODS

In this section, we detail the methods used in our study of laughter for 1) obtaining laughter / non-laughter candidates for annotation, 2) laughter annotation, 3) the study design (ie. assignment of laughter candidates to annotators, and related decisions) and 4) automatic laughter assessment.

7.5.1 LAUGHTER CANDIDATE GENERATION

To obtain a set of laughter candidate segments (thin slices) to be annotated in our human study, the authors localized any *possible occurrences* of laughter in the dataset by watching the audiovisual recordings for every data subject and segmenting perceived laughter episodes using the annotation software ELAN [120] by indicating the start and end of each laugh on top of the audio waveform. We were deliberately inclusive by annotating segments when in doubt. Annotations closer than 1s apart were considered a single laughter episode. Not all segments, however, were visible in the videos. Therefore, we additionally annotated the cameras, if any, in which a particular laughter episode was visible. Segments present in multiple cameras were only considered once by randomly picking one of the cameras. Segments not present in the video were discarded.

NEGATIVE CANDIDATE GENERATION

As negative samples, we extracted a number of segments likely containing no laughter from the rest of the dataset. To avoid having mostly segments of *listening behavior* in this negative set (our conversing groups were often large), we sampled negative candidates from speech utterances as given by our VAD labels. Additionally, since some data subjects were much more likely to laugh than others, we sampled the distribution of negative samples per subject proportional to the distribution of positive samples. Concretely, our sampling procedure is:

1. samples a subject S with probability $P_L(S)$ where S is the probability of a positive laughter candidate belonging to subject S .
2. samples a speech utterance uniformly from the set of speech utterances of S of length $l_{min} < l < l_{max}$ where l_{min} and l_{max} are the lengths of the shortest and longest laughter episodes. This was done to avoid very long speech utterances from being introduced as negative examples.

EXPANDING CANDIDATES IN TIME

Finally, laughter candidates (positive and negative) were expanded in time. The goal was to more closely resemble the process of annotating laughter in the wild, where it is unknown when laughter might happen, and allow the annotator to understand some of the context of the scene. To this end, we expanded each segment at both ends with a duration randomly (uniformly) sampled between 1.5s and 3.5s. We set the bottom of this range (1.5s) to be close to the mean length of a laughter episode. Empirically, this was enough to process the scene and be ready for annotation. We set the top of the range (3.5s) to obtain total segment lengths below 10s to maintain the annotation process fast. We used a uniform distribution to minimize the predictability of the location of the laughter episode.

4. Mingling dataset (side-elevated videos, individual audio + body acceleration)

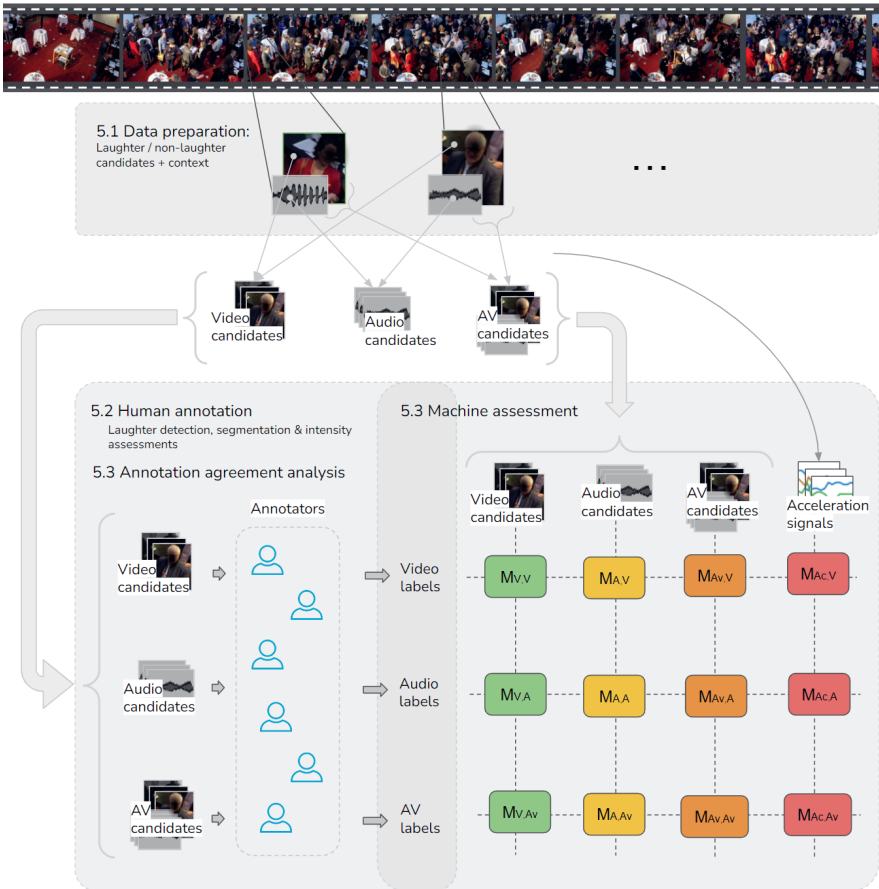
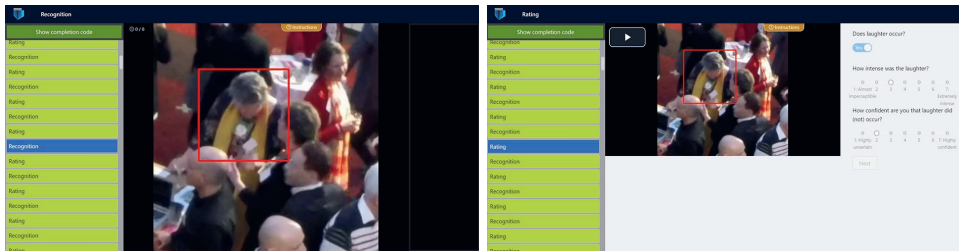


Figure 7.2: Overview of our study. From a mingling dataset with video, individual audio, and accelerometer readings (Section 7.4), we extracted pre-annotated segments of potential laughter and speech, each of 7s in length. These segments were annotated for laughter presence, segmentation, and intensity under three conditions: audio-only, video-only, and audiovisual. We analyze the labels directly (Section 7.6.1) and use the different sets of labels to train and benchmark models for laughter detection, segmentation, and intensity estimation (Section 7.6.2).

SPATIAL LOCALIZATION VIA BOUNDING BOX

Since our side-elevated camera views captured most of the interaction scene, the target subject needed to be extracted or indicated to annotators. This was done by annotating a single, tight, bounding box around the target person for the first frame of the video. To allow annotators to use the visual context of the scene while providing good visibility of the target subject, videos were cropped beyond the borders of this bounding box by multiplying its width and height by 3 (constrained to fit within the frame) and maintaining its center. Our observations showed that this was in most cases enough to capture the interlocutors of the target person. The target person’s box was shown to the annotators before the start of the video (see Figure 7.3a).



(a) Recognition and continuous annotation step.

(b) Intensity and confidence annotation.

Figure 7.3: Screenshots of the annotation interface in Covfee [110]. The two steps shown were repeated for every example that an annotator rated. In (a) annotators were shown a target person marked by a red box, and part of the scene around the target, and instructed to hold down a key when laughter was perceived to be occurring by the target person. The interface provided visual feedback when the key was held down. In (b), subsequently, annotators rated laughter intensity (Likert scale 1-7) and their confidence in their assessment (Likert scale 1-7).

7.5.2 ANNOTATION OF LAUGHTER CANDIDATES

Central to our study of laughter annotation is the design of the process to be followed by annotators. The first step in the annotation of laughter in the wild is the localization of the laughter episodes in time.

Actions are traditionally localized in videos using tools such as ELAN [120], where the user localizes the start and end frame of the action by drawing an interval on top of an audio waveform. In tools such as Vatic and CVAT, actions are annotated via flags, which are turned on for the frame when the action is deemed to start, and off for the end of the action. In affective computing, *continuous annotation* techniques are commonly used to annotate variables such as arousal and valence. In *continuous annotation*, annotators control the value of the target variable while watching the subject in video, usually without pause. This has the advantage of letting the annotator perceive the behavior without interruption and being efficient and predictable in terms of time needed to annotate. On the other hand, continuous methods also necessarily introduce a reaction time delay. Multiple techniques have been proposed to mitigate these delays.

We chose to make use of continuous annotation for our study due to the mentioned advantages. We mitigated annotation delay by making use of an experimentally defined offset, as detailed in Section 7.5.2. We made use of a binary action localization technique implemented in the Covfee framework [110], which asks annotators to hold down a

keyboard key when they perceive laughter to be occurring. Its graphical interface is shown in Figure 7.3a. This process allows annotators to maintain focus on the videos by minimizing the input effort, while still giving us access to high-resolution segmentation of laughter. Since the annotation time is shortened and predictable, this process also allowed us to obtain more annotations per annotator, relevant to our study design (Section 7.5.2).

After the continuous annotation step, for each candidate, we asked annotators explicitly whether they perceived laughter to occur, their perceived laughter intensity, and their confidence in their laughter ratings (Figure 7.3b). Annotators could replay the laughter episode if they desired.

CROWD-SOURCED ANNOTATION PROCESS

As introduced in Section 7.5, answering our research questions requires annotations of laughter under three conditions: audio-only, video-only, and audiovisual. Measuring agreement within a condition imposes the requirement that at least two annotators rate each (*candidate, condition*) combination. A sufficient number of candidates must also be annotated to be able to train our computational models and measure agreement over a large enough set. Finally, for access to a large pool of annotators, annotations would be crowd-sourced and each annotation HIT (human intelligence task) should ideally not last longer than approximately 45 minutes to avoid fatigue. In our tests, we estimated each candidate to require about 30 seconds for annotation. This imposes an upper bound on the number of samples per annotator of around 90, which we reduced to 84 to have room for error.

One other natural choice to consider was whether to use a between-subjects or within-subjects design. To maximize the number of annotators per condition, we opted for a study where each annotator takes part in all three conditions. To avoid bias, we impose the restriction that one annotator never annotates the same candidate under different (or the same) conditions, ie. one annotator rates three disjoint sets of candidates.

According to these design decisions, we divided our 659 candidates into 7 sets of 84 examples and one set containing the remaining 71 candidates. Each of these candidate sets was in turn divided into three equal-size subsets (for the three conditions). Each permutation of these three subsets resulted in a different human intelligence task (HIT), each containing the same candidate subsets but mapped to different conditions. Figure 7.4 is a diagram of this process for each set of 84 candidates. Each HIT was completed by two annotators. Annotating all candidates required 48 annotators in total. This design allowed us to compute pair-wise inter-annotator agreement (per condition) over sets of 28 paired ratings, for 24 distinct pairs of annotators.

ANNOTATION HITS

We crowd-sourced our annotations to 48 annotators via the Prolific crowd-sourcing platform [378]. We implemented the complete annotation flow using the Covfee annotation framework [110]. Each HIT contained several introductory tasks and examples, followed by three annotation blocks, one for each modality condition. The order of video-only, audio-only, and audiovisual blocks was randomized to avoid ordering bias due to factors like fatigue. The ordering of laughter examples within each block was also randomized for the same reason. The detailed structure of a HIT is presented in Appendix 7.A.1. Statistics

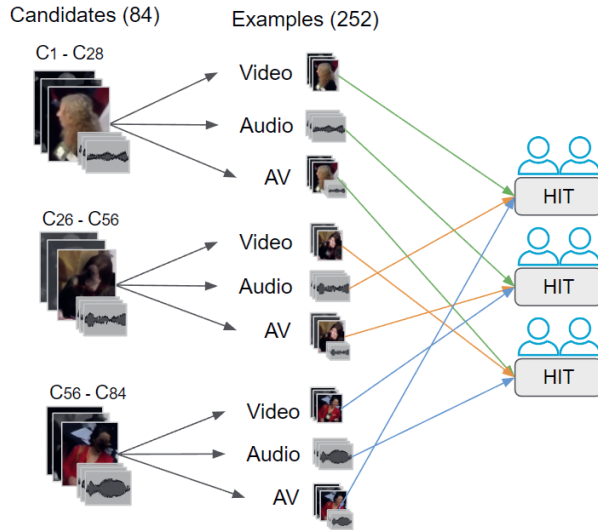


Figure 7.4: Structure of the annotation stage of our study. Sets of 84 randomly-selected candidates are separated into 3 equal-size sets of 28 candidates. Candidates are then separated into their audio and audiovisual modalities and assigned to HITs such that each HIT contains 28 distinct candidates per condition. Each HIT was annotated by 2 annotators.

7

of the ratings provided by each annotator, time to complete the experiment, and experience ratings are presented in Appendix 7.A.2.

ANNOTATION DELAY CORRECTION

Delays in continuous annotation have been investigated within the affective computing community for continuous-value annotations of affective dimensions. Some works have proposed machine learning models that are robust to annotation delays [118, 119]. Mariooryad et al. [117] proposed a method for correcting delay by maximizing mutual information between the continuous label time series and an auxiliary signal containing facial keypoints. However, the authors also showed that simply offsetting annotations by a constant value resulted in performance comparable to that of more complex schemes.

Despite these results, it is unclear to what extent delay depends on the particular actions being annotated. We therefore decided to measure delay directly for our task and annotators. At six points in each annotation HIT (two per condition, see Section 7.5.2), we inserted special *calibration* (positive) laughter examples, which were the same for all annotators. We precisely labeled the onset and offset times of laughter in these six examples, using ELAN [120]. This allowed us to calculate a delay in the annotator’s continuous labels, to approximate the average delay of each annotator. We used this average annotator delay as a correction offset for an annotator’s labels.

7.5.3 MEASURING INTER-ANNOTATOR AGREEMENT

We designed our study for the computation of inter-rater agreement, or reliability, within and across conditions. Cohen's Kappa, Fleiss' Kappa, and Krippendorfs Alpha are some commonly used measures of agreement. For nominal values (eg. laughter / non-laughter) Cohen's Kappa is capable of computing agreement between exactly two annotators. Although Cohen's Kappa is subject to biases in some instances, it still has been recommended by previous work for fully crossed designs with multiple coders, by computing the average of pairwise agreement [379]. Since each of our annotator groups rated a set of examples not rated by any other pair (ie. our study consists of a set of fully-crossed designs), we used this approach to measure agreement for nominal values.

Cohen's Kappa is however not appropriate for interval / ordinal values like laughter intensity (Likert scale 1-7). Here, we used Krippendorff's alpha, a reliability measure applicable to any number of raters and which adjusts for small sample sizes. We averaged pairwise Krippendorff's alpha values over rater pairs.

7.5.4 AUTOMATIC, LAUGHTER DETECTION, INTENSITY ESTIMATION, AND SEGMENTATION

Video-based models for detecting, assessing (eg. intensity) and segmenting actions have been extensively studied in computer vision and pattern recognition (Section 7.2.2). We make use of modern approaches within these fields. Regarding the video modality, due to the small size of our dataset, training state-of-the-art methods from scratch would be infeasible. We focused on approaches with pre-trained models available to use as feature extractors. Among those, 3D convolutional neural networks (CNNs) are known to reliably achieve top performances in action recognition benchmarks. We decided to make use of a 3D ResNet pre-trained on Kinetics-400, a large action recognition dataset with 400 action classes and over 300000 labeled video clips. The network implementation and models are available as part of the *Pytorchvideo* library [323].

Regarding audio-based models, work by Gillick et al. [358] investigated laughter detection in two datasets with significant background noise. One of these, the Audioset dataset [380] is freely available to download. This dataset of 10-second clips from Youtube videos recorded in a variety of in-the-wild settings contains 5696 clips labeled as containing laughter. In their implementation, the authors provided a list of randomly sampled no-laughter clips to complete the dataset with negative. Given that this dataset had more examples and a variety of subjects than ours, we decided to pre-train the audio-based model on it. We made use of the same model proposed by Gillick et al. [358]: a 2D ResNet model operating on the spectrogram of the audio inputs. We trained on 85% of the dataset, with 15% separated to determine a good stopping point. We otherwise used the same hyper-parameters used by the authors.

As motivated in section 7.3, we made use of acceleration as an additional modality capturing body movements. As an acceleration-based model, we made use of a ResNet variant for time series, implemented as part of the *tsai* library [306]. Given the much lower dimensionality of the acceleration data (compared to video and audio), and the lack of availability of comparable acceleration datasets, we trained this model from scratch.

For both video and audio models, we used pre-trained models as feature extractors by freezing all parameters and removing network heads. For classification, the features

output by the base networks (with dimensionalities: 2304 (audio), 8192 (video), and 128 (acceleration)) are fed into a head consisting of a linear layer followed by an output sigmoid layer and binary cross-entropy loss, standard choices for binary classification. For multimodal evaluation, we concatenate the features from multiple models before the head of the network.

We decided to approach intensity estimation as a regression task, given the interval nature of the ratings. We follow the same model structure, but we removed the sigmoid computation from the output and used L2, or mean squared error (MSE) loss, a standard choice for data with no outliers.

For segmentation, we decided to approach the task as the estimation of a binary mask (ie. of our continuous binary annotations). This would allow us to use the same base networks and pre-trained models. However, multimodal fusion should now be done earlier, since the time dimension encodes information likely useful for segmentation. We therefore implemented separate segmentation heads per modality, which are fused at the output via average pooling. For all models, we apply pooling and convolution operations over the spatial and channel dimensions and up-sample the time dimension to the length of the target segmentation mask (45). Details of the architecture are presented in Appendix 7.B

GENERATING TRAIN AND TEST SAMPLES FROM LAUGHTER ANNOTATIONS

Given that the examples seen by laughter annotators contained a significant amount of context, using the complete 7s candidates for the machine learning tasks would not be ideal given the much shorter average duration of laughter. Furthermore, our models made use of fixed-size inputs, and the examples rated by annotators were not fixed in length. To address the situation, we used the continuous binary labeling signal as a reference, and sampled shorter positive windows around its positive sections (ie. exactly where laughter was detected to have occurred). Figure 7.5 shows a simplified depiction of the process. Given a binary annotation signal with at least one positive segment, we consider the intervals within its positive segments as candidate window locations. We sample uniformly from these locations to select the window center, which determines the limits of the window. For negative examples (ie. with no positive segments), we consider every location in the signal to be a candidate for the window center (ie. we perform a random crop).

To determine the size of the window, we looked at the distribution of laughter lengths, as obtained from our continuous annotations. The average laughter length was 1.14s, with a long-tailed distribution such that 80 percent of laughs were under 1.56s. We chose a length of 1.5s as this length guarantees that most laughter segments will be contained in the window without excessive non-laughter context.

In evaluation, to avoid randomness, instead of the sampling procedure the window is always centered on a positive segment for positive examples. For negatives, the window is always in the middle of the complete candidate.

We followed the same process for the three tasks of laughter detection, intensity estimation, and segmentation, but the labels differ per task. For detection, the sample is labeled positive when it comes from a positive annotation segment, and negative otherwise. For intensity estimation, the segment is labeled with the intensity label (Likert scale 1-7) for the laughter candidate. Negative samples were included and assigned an intensity of zero. For segmentation, the target is a vector corresponding to the continuous binary annotations (30 *fps*) within the target window (vector of size 45 for our 1.5s windows).



Figure 7.5: Illustration of the process used to select positive laughter samples for our machine learning tasks. Given the binary laughter / non-laughter annotations for a particular segment, we select a location for the window center from the positively-annotated intervals in the signal. We then extract a window of 1.5 seconds around the chosen center. We pad if necessary.

Note that our annotation study involved two raters per candidate and condition. Both of these continuous ratings are included in the sampling process for each epoch.

EVALUATION PROCEDURE

For evaluation, we made use of standard metrics for each task. For classification, we make use of the area under the ROC curve (AUC), a metric designed for binary classification and invariant to class imbalance. For regression, we make use of Mean Squared Error (MSE). We also make use of AUC for segmentation, where we treat every window element as one separate prediction. Although metrics like Intersection over Union (IoU) are more commonplace in segmentation, we made use of AUC due to it not being affected by class imbalance.

We evaluated via 10-fold cross-validation, to obtain an aggregated performance measure over the whole dataset. We used the first fold for tuning the number of epochs to train for (per combination of modalities) and excluded the first fold from the evaluation.

7.6 RESULTS

7.6.1 COMPARISON OF HUMAN LAUGHTER ANNOTATION AGREEMENT ACROSS MODALITIES

To test our hypotheses around differences in annotations across modalities, we started by calculating inter-annotator agreement within and across modalities via pairwise computation of agreement metrics (Section 7.5.3). Tables 7.1a and 7.1b show the results of our agreement calculations for laughter detection and intensity rating. Note that within-modality calculations are averages over 24 (pairwise) comparisons and between-modality calculations are averages over 96 pairs. Standard deviations are shown in parentheses (calculated across pairs). Agreement scores for laughter detection (Table 7.1a) show that the audio and audiovisual conditions have a greater agreement between them, with video

Table 7.1: Precision, recall, and inter-annotator agreement measures across modalities.

(a) Laughter detection inter-rater agreement (Cohen's Kappa)

Condition	Audio-only	Video-only	Audiovisual
Audio-only	0.823 (0.153)		
Video-only	0.396 (0.186)	0.550 (0.146)	
Audiovisual	0.795 (0.144)	0.424 (0.183)	0.805 (0.144)

(b) Laughter intensity inter-rater agreement (Krippendorff's alpha)

Condition	Audio-only	Video-only	Audiovisual
Audio-only	0.664 (0.162)		
Video-only	0.237 (0.228)	0.394 (0.279)	
Audiovisual	0.663 (0.168)	0.267 (0.239)	0.697 (0.165)

being significantly lower. The video condition had higher within-condition agreement than agreement with other modalities.

Agreement in intensity estimation (Table 7.1b) shows a similar trend. The lowest agreements, once again, were found between audio and video (0.396 ± 0.186) and between audiovisual and video conditions (0.424 ± 0.183). These are lower than all within-modality agreement scores, even that of video. This suggests that the *concept* of laughter intensity was perceived differently when audio was available and when it was not. Note that agreement in laughter intensity was only calculated between examples labeled positively (as laughter) so that scores are not biased by detection ratings.

We tested the effect of the annotation condition on intensity ratings via a linear mixed effects model with the condition as a fixed effect. The annotator ID was used as a grouping variable (random effect) to control for annotator-specific variance. We fitted the model only on the subset of positive laughter annotations. We found the condition to have a significant effect on intensity ($p = 0.00223$). A cluster bootstrap analysis revealed that laughter was annotated as being significantly less intense in audiovisual (95% confidence interval of $[-0.44, -0.0406]$) and video-only conditions (95% CI of $[-0.45, -0.0482]$). This is a relatively small effect considering the scale of our intensity ratings (1-7).

To get further clarity about the quality of video-based annotations, we compared them to *reference annotations* from the audiovisual condition. We consider the audiovisual condition to be the most ideal one due to annotators having access to both modalities. However, laughter is not always a clear signal and therefore we consider this to be a *reference* set rather than ground truth. We derived this set of binary labels via majority voting, for each (*candidate, condition*) pair, on the annotator ratings (2), and the expert rating (1), for a total of three votes. We used this reference set for calculation of precision and recall scores.

Table 7.2 shows the precision and recall scores for the three annotation modalities w.r.t. the reference annotations. Results show that false positives are rare in our annotations. Recall scores show more differences, with video being lower than both audio and audiovisual scores. This aligns with our hypothesis that the video modality is not enough to detect many

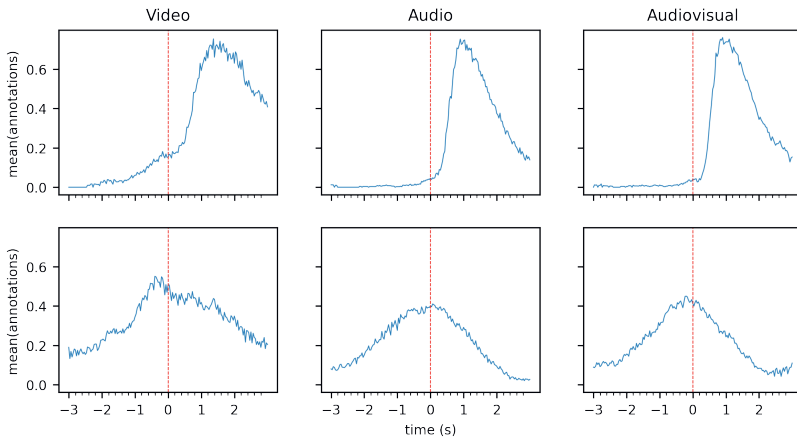


Figure 7.6: Aggregated onsets and offsets w.r.t. reference annotations from different modalities.

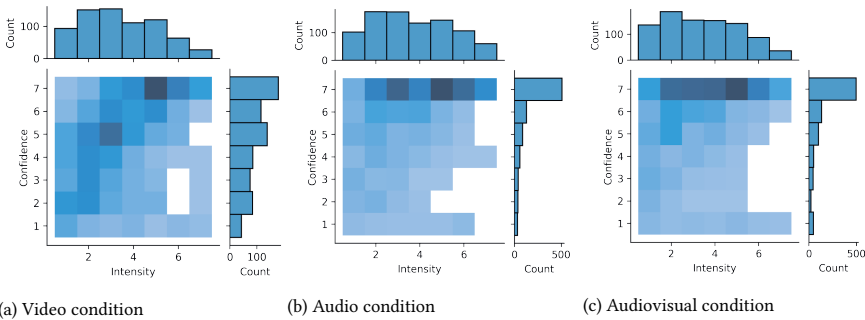


Figure 7.7: Joint distribution of confidence and intensity values. Both were annotated using a Likert scale (1 – 7). Confidence indicates the confidence of the annotators on their laughter annotation for the candidate segment.

Table 7.2: Precision and recall w.r.t. to annotation reference.

	Audio-only	Video-only	Audiovisual
Precision	0.9645	0.8915	0.9812
Recall	0.9405	0.7024	0.9578

episodes of laughter (ie. a large number of false negatives). As expected, the audiovisual condition had the highest precision and recall. Note however that reference annotations were obtained from audiovisual labels, and this might cause the numbers to be artificially inflated.

Comparing agreement in localization of laughter is less straightforward since multiple variables are involved. We decided to do so qualitatively, by plotting the mean value of annotations, across different examples, around reference onsets (rising edge of the binary signal) and offsets (falling edge). Ideally, annotators would agree exactly on the onset of the laugh and we would observe a step-like plot. In practice, onsets and offsets vary per annotation and a curve is observed. Figure 7.6 shows the mean value of annotations around onsets (key pressed) and offsets (key released). These are aggregated over different laughter samples and show once again better agreement when audio is present. Offsets display less agreement (flatter shape) than onsets. We attribute this to the end of a laugh is often less clear than its start, blending in with speech or other utterances.

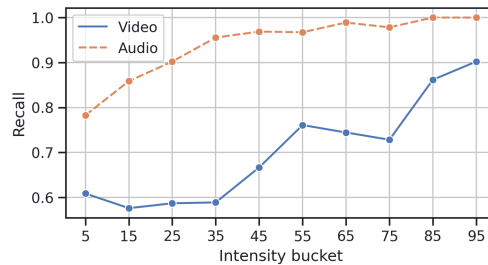
We complete our analysis by looking at annotator confidence, as an indication of the difficulty of the task in each modality. Figure 7.7 we plot the distribution of laughter intensity and confidence values for the three conditions. We used a Likert scale for both of these ratings, and the distributions are therefore discrete. While intensity distributions are similar across the three conditions, the confidence histograms make clear how much more challenging the video-only condition was to annotators. The wider distribution reveals a clear correlation between laughter intensity and confidence in their annotation, as would be expected.

THE ROLE OF LAUGHTER INTENSITY

The results in section 7.6.1 showed that video-only laughter annotations have lower recall than audio-only annotations. We hypothesized, however, that this is likely due to the difficulty of detecting low-intensity laughs, which are likely to have less salient associated body movements.

To verify this, we separated our dataset by laughter intensity. We obtained a single consolidated audiovisual intensity rating per example by averaging the intensity ratings from both annotators. We then separated the dataset into 10 intensity buckets, from lowest to highest intensity. To ensure a sufficient number of samples per bucket, we used percentiles to define the bucket sizes, such that bucket i includes laughs between the $(i \times 10)$ th and $((i + 1) \times 10)$ th percentiles of intensity. We computed recall for each bucket. Figure 7.8 plots the results of this analysis. As expected, recall of both audio and video conditions increases with the audiovisual intensity of the laugh. As hypothesized, video recall tends to approach audio recall for the most intense laughs. It stands out, however, that the gap between them never closes completely, even for the 10% most intense laughs.

Figure 7.8: Laughter recall against the (audiovisual) intensity of the laughs. The x axis indicates the middle of the percentile bucket (eg. 15 is the bucket with laughs between the 10 th and 20 th percentile). As intensity increases the recall of video-only annotations approaches that of audio-only annotations.



This can be understood in the light of the findings of Section 7.6.1, where it was shown that intensity ratings in the audio and audiovisual have high agreement, but they both have low agreement with the video-only ratings. Our consolidated audiovisual intensity ratings, therefore, do not reflect intensity as perceived in the video-only condition.

7.6.2 EFFECT OF LABELING MODALITY ON SUPERVISED LAUGHTER TASKS

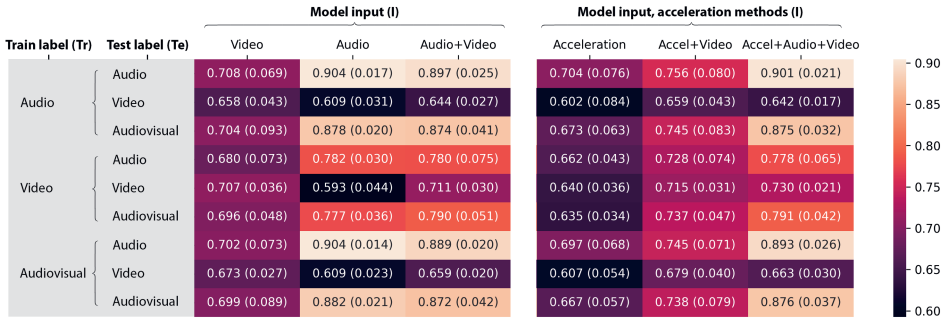
Although the analysis of inter-annotator agreement performed in the previous section is relevant to understanding differences in labels themselves, it does not ultimately answer the question of how useful annotations acquired from different modalities are for training automated models.

The answer to this question is nuanced. We might have access to video-based annotations of laughter and want to understand if training a video-based action recognition model with them would help detect vocalizations of laughter. However, asking the reverse question is also of interest: would audio-based annotations result in a model capable of detecting the characteristic body movements of laughter? Furthermore, would audio-based annotations be the most appropriate, or would it be preferable to label the same modality that is input to the model?

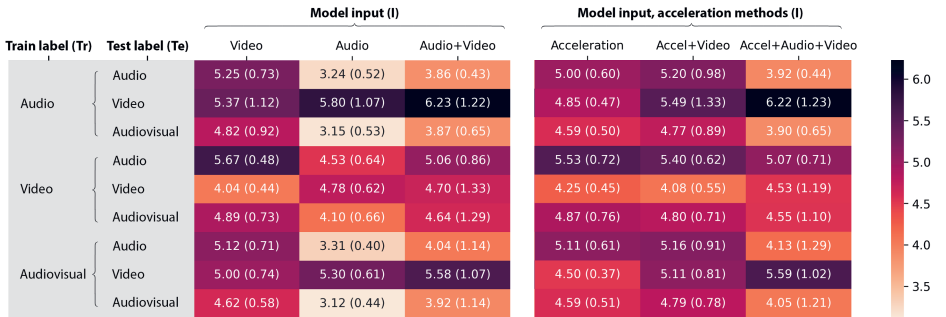
The goal of this section is to investigate the impact of annotation modality on trained model performance. Machine learning methods can naturally accept different modalities of input data and we are interested in the relationship and possible interactions between input modality, training label modality, and testing label modality.

To this end, in line with the tasks that annotators performed in our human study, we trained and evaluated models for the tasks of laughter detection, intensity estimation, and segmentation (Section 7.5.4). For each of these tasks, we evaluated models for all possible combinations of six different input types (acceleration, audio-only, video-only, video+acceleration, audio+video, audiovisual), training label modalities (audio, video, audiovisual) and testing label modalities (audio, video, audiovisual). We used acceleration as an additional input to leverage the wearable data available in our dataset. Wearable acceleration has been found in previous work to be a useful proxy for body movement. Positive and negative examples were generated for our experiments from the human

(a) Classification into laughter / non-laughter (AUC, higher is better).



(b) Regression of laughter intensity (MSE, lower is better).



(c) Laughter segmentation (AUC, higher is better).

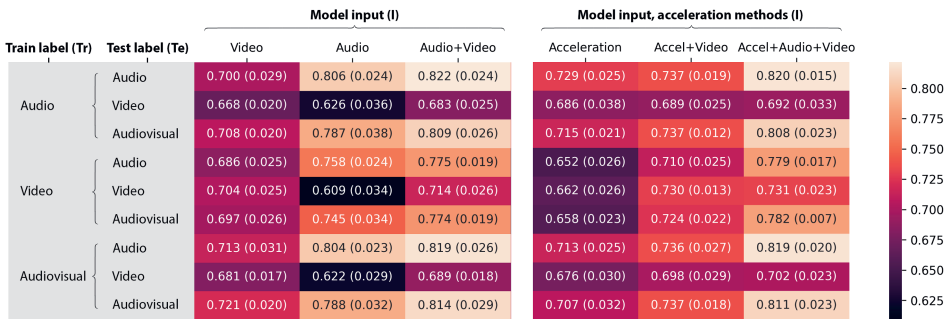


Figure 7.9: Results of our machine learning experiments (10-fold cross-validation). Columns correspond to different model input modalities. Rows correspond to training label modality and testing label modality. For example, *Audio* > *Video* indicates a model trained with labels acquired from audio alone, and tested on labels acquired from video alone.

laughter annotations per the procedure in Section 7.5.4. We evaluated each model using 10-fold cross-validation and the Area under the ROC Curve (AUC) as evaluation metric, as explained in section 7.5.4.

Figure 7.9 presents the results of our machine learning runs. For readers' convenience, we may refer to the results in the tables using the abbreviations in the column labels. For example, $I = Acceleration, Tr = Video, Te = Video$ localizes the fourth cell in the *Acceleration* column.

It is clear that for all tasks (audio-)visual inputs trained (audio-)visual labels ($I = Audio|Audio + Video, Tr = Audio|Audiovisual$) had the best performances, except when applied to video-based labels ($Te = Video$). This is likely explained by these methods detecting many positives that are not labeled in video, due to having low body movement intensity. In defense of video-based labeling, it stands out that models with video inputs show no significant differences in performance across training and testing label modalities ($I = Video$). In other words, the modality used for labeling had no effect on the final performance of video models. The acceleration and video+acceleration methods had a similar behavior, with no significant differences due to training label modality. This provides some support for the use of video labeling for model inputs capturing body movement information. Furthermore, video labels were enough for training an audio-based detection method with an AUC of 0.782 ($I = Audio, Tr = Video, Te = Audio$), a performance drop of less than 0.15 AUC with respect to audio labels.

Note that classification results (Figure 7.9a) display a pattern similar to that of segmentation (Figure 7.9c). For segmentation, however, scores of audio-based methods ($I = Audio$) are lower than for classification, while the scores of video-based methods ($I = Video$) remain the same, making the video-based methods more competitive with the audio-based ones, though still significantly worse-performing for most label combinations. Regarding the acceleration modality, it stands out that *Acceleration + Video* methods often improved over both modalities in isolation, supporting the idea that these modalities are complimentary.

The results of intensity regression methods (Figure 7.9b) are more particular. In contrast to classification and segmentation, most multimodal models performed worse (higher MSE) than audio-only models for the same labels (ie. $I = Audio$ generally has the lowest MSE), meaning that adding input modalities tended to affect the model. We also observe that video and acceleration regression models perform best when trained and tested on video labels, but training on audio and testing on video or vice-versa results in some of the worst performances. This aligns with the findings from the annotation experiments that the intensity of laughter in the video and audio modalities are incongruent.

7.7 DISCUSSION

Our inter-rater agreement results present evidence that annotation of laughter occurrence, intensity, and temporal extent can differ substantially across annotation modalities. Per our hypothesis, video annotations had lower agreement than audio and audiovisual ones. When comparing against audiovisual reference annotations, we found recall to be worse in the video condition. Differences in precision scores were lower, with all modalities being close to the 90% to 95% range. These findings suggest that video-based annotation of laughter, while feasible, should not be used in applications requiring high recall. Zooming into the issue of low recall revealed that recall improves for video annotations the more

intense the laughs being considered, likely as a result of higher saliency of body movement cues. In the light of previous work [22], this means that video-based laughter annotations are more likely to capture humorous laughter, strongly associated to high intensities, than the more common rule-bound conversational laughter.

Regarding differences between audio and audiovisual conditions, our results revealed high within and between-condition agreement (0.8 for detection, 0.66 for intensity estimation) between them. These results validate the use of audio as the primary modality for laughter annotation, but they are not without nuance. Although they indicate that there was a clearer shared concept being annotated when audio was present, video annotations had higher within-condition agreement than agreement with audio and audiovisual annotations. This suggests that there is a different concept being perceived in the video condition with some consistency. Given the low recall of the video condition, we interpret this to indicate that false negatives (w.r.t. audiovisual reference) are missed systematically, likely due to the absence or subtlety of their visual cues. Systematic false positives across annotators also likely contribute to these results, though to a smaller degree. In other words, there appears to be incongruence in the perception of laughter occurrence across modalities.

These results set the stage for the question explored in our machine learning analysis: is perception of laughter in the visual modality a meaningful concept to annotate for the purpose of building detectors, despite its incongruence with audiovisual laughter?

Importantly, we measured a similar incongruence in laughter intensity ratings, where only positively labeled segments were included in the agreement calculations, indicating that laughter intensity is not perceived in the same way when audio is present and when it is not. Such incongruences in laughter intensity across modalities have only been studied in the context of laughter synthesis. Niewiadomski et al. found that laughter episodes with incongruent body movement and vocalization intensities were rated as less believable [34]. This would seem to go against our results, which suggest that significant incongruence exists in in-the-wild laughter perception. However, the magnitude of the incongruencies used (which can be controlled in a synthesis study, but not in the wild) could explain this discrepancy.

Our results have implications in studies of laughter intensity [5, 34, 106, 348–351], suggesting that the concept of laughter intensity should not be treated as a scalar property of the laughter episode, but rather as a nuanced evaluation affected especially by the modalities available to the observer. In particular, the question of whether a clear distinction should be made between the intensity of body movements and the intensity of the sound of laughter deserves consideration. McKeown et al. already asked the question of whether laughter body movement intensity itself should be considered multi-dimensional [106], but the distinction between visual and auditory intensity has not been considered before, to the best of our knowledge.

Our findings lead us to the fundamental question of what is laughter intensity in the wild. Are the observed differences across modalities mainly a product of imperfect recording conditions, or would we observe them too under ideal conditions? (eg. in face-to-face interactions). While in our dataset subjects prioritized audio in the multimodal condition, it is not clear if body movement information would be prioritized in other datasets in which it is easier to perceive (ie. with consistent access to the face or upper body), or in which

the audio is harder to perceive. We consider it likely that in such cases visual information will play a more important role, but more work is necessary to provide an answer to these questions.

Despite the lower inter-annotator agreement in the video condition, our machine learning experiments with different combinations of model inputs, training label modalities, and testing label modalities, revealed that model performance was the same across labels for models trained using video and acceleration inputs, both of which capture body movements. This was regardless of the evaluation modality. In other words, annotating laughter (traditionally understood primarily as a vocalization) from video alone may be perfectly valid when the goal is to optimize model performance. We think that the reason for such results is explained by our human annotation analysis. Concretely, episodes with lower intensity were most commonly missed (w.r.t. to the audiovisual reference). The subtlety of these training samples would presumably make them more challenging for the learning algorithm, and therefore their absence would not have an adverse effect on performance. We obtained these results in a challenging dataset, where many positive (audiovisual) laughter episodes were missed by annotators, and used a modern action recognition 3D-CNN. It is once again unclear whether these results would translate to a dataset with more consistent access to, for example, facial visual information. The presence of visual cues could improve the model, but their subtlety could be a challenge to most state-of-the-art models. More work is warranted in this direction.

Our results provide validation for previous works using video-only labeling to train laughter assessment models from body movements [103, 348], and datasets providing video-only annotations [89]. Recording audio is not only a technical challenge (especially for large groups), but the use of video labeling is also more privacy-conscious as it avoids the need for recording the content of conversations. However, the fact that annotations obtained from video are largely incongruent with audiovisual annotations should be a consideration in many studies.

We think that these results could have wider implications if they generalize to other multi-modal social signals with manifestations in body movement. Speaking status (or voice activity) and back channels have been of interest in previous work [227, 381]. Video-only annotations of speaking status have been used in previous work [81, 89, 95], but the implications in model performance of this annotation choice have not been explored. Our results would suggest that it is possible to annotate speaking from video alone without an adverse effect on the model's ability to detect speech, but further work is necessary to provide validation for other multimodal social signals besides laughter.

7.7.1 LIMITATIONS

We consider the main limitation of our work to be that we used only one dataset in our experiments. Our dataset is however representative of one of the most challenging scenarios for the perception of laughter from video: with little access to the face of the participants, different views and distances to the camera, low light conditions, and significant occlusion of parts of the body from other participants in the scene. We therefore considered it a useful data point to study. We expect that more traditional front-facing datasets with consistent access to the body and face of the subjects will result in lower differences in agreement and model performance between the video condition and the audio and audiovisual ones. We

think clear access to the face may negate the incongruence observed in laughter intensity ratings, since facial features may share more information with the laughter vocalization than overall body movement does.

ACKNOWLEDGEMENTS

This research is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606. We acknowledge our crowd-sourced annotators who very conscientiously participated in our preliminary tests and final study, and in some cases provided valuable voluntary feedback.

Appendices

Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in the wild

7.A ANNOTATION EXPERIMENT DETAILS

7.A.1 HIT STRUCTURE

The HITs in our human annotation study consisted in a sequence of pages or tasks to be followed by an annotator in order. In general, each HIT contained several introductory tasks and examples, followed by three annotation blocks, one for each modality condition. The order of these three blocks, and of the examples within was randomized for each instance.

In detail, each HIT consists of the following tasks:

1. Consent Form (5 min). Participants were asked to agree to an End User License Agreement required to access our dataset. This was put in place to protect the privacy of data subjects and were required to continue with the annotation.
2. General instructions (5 min). Introduction to the HIT, informing the annotator about the different sections/conditions, the need for audio equipment, and the structure of the HIT.
3. Reaction time test explainer (2 min). An example reaction time test, with instructions to let the annotators familiarize themselves with the test.
4. Example laughter segments (3 min). Three example segments where the annotator was asked to continuously annotate and then rate laughter segments. We chose segments where laughter was clear and evident since the sole purpose of these segments was to let the subjects become familiar with the process.
5. Video-only block (28 segments, 10 min). One video per page. The participant must play the video and press a keyboard key when they perceive laughter to be occurring. At the end of each segment, the participant must provide a rating of laughter intensity using a slider with a continuous range between 0 and 10 and a rating of confidence in their laughter annotation. This is all explained on a page with instructions at the start of the block.
6. Audio-only block (28 segments, 10 min). Same as above. Participants must play a web video containing no image (only audio) and similarly press a key when they think they can hear laughter. In the instructions page, participants are asked to test their audio equipment (speakers, headphones).
7. Audiovisual block (28 segments, 10 min). Same as above, but now annotators get access to both audio and video.
8. Optional feedback (1 min). Annotators were asked to (optionally) rate their experience and give free text feedback on the process.

7.A.2 ANNOTATION EXPERIMENT STATISTICS

In this section, we present statistics from our annotation experiments from 48 annotators, including the number of times the annotator detected laughter, the distribution of intensity and confidence ratings, and the time taken to complete the experiment. Additionally, we requested an optional rating of their experience completing the HIT (*How would you rate your experience in completing this experiment?*) on a scale from 1 to 5.

Table 7.3 shows the annotation details, with each row being one hit/annotator. G indicates the HIT group. HITs within the same group contain the same laughter/non-laughter samples. N indicates the HIT number within the group. HITs with the same N are identical, except for the (random) ordering of the samples within each condition. HITs with different N contain the same samples but are assigned to different conditions. Each row corresponds to one annotator (HIT). $\# \text{ positive}$ indicates the number of times laughter was detected by this person. Intensities indicates the histogram of laughter intensities by each annotator (positive examples only). Each number (in order) corresponds to one step in the Likert scale (1-7). Note that per our pre-annotations of laughter, 59 of the 84 examples in each HIT contained laughter, while the rest only contained speech. Note also that time taken in completing the experiment was measured as the difference between timestamps in the data sent at the beginning and end of the experiment, and does not contemplate the fact that annotators could have taken breaks in between.

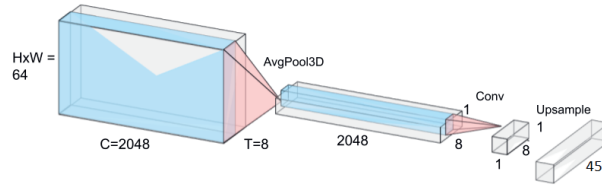
We also allowed annotators to provide free text feedback (*Do you have any comments about the process? Did you find it frustrating, tiring, or too long? Were the instructions clear? Did you have any issues with the tool?*) about the process. Here, most annotators who answered reported surprise at the originality of the task, some saying they found it interesting and/or they had never completed an experiment of this kind. Some commented about the instructions, reporting them to be clear. Some annotators reported the experiment being long/tiring.

Table 7.3: Statistics of the annotation HITS. See Section 7.A.2 for column name details.

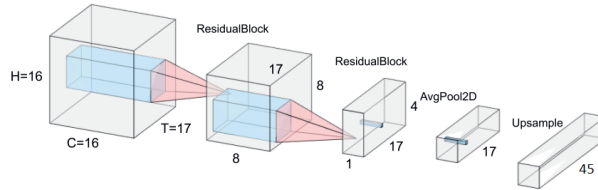
ID	G	N	# positive	Intensity	Confidence	Time taken	Rating
1	0	1	53/84	13-11-11-09-06-03-00	03-09-07-07-08-15-35	43.0	4/5
2	0	1	55/84	17-19-09-05-04-02-00	02-02-08-06-06-13-47	45.5	4/5
3	0	2	49/84	10-11-12-04-08-02-02	03-04-09-03-15-16-27	41.6	4/5
4	0	2	61/84	08-10-10-01-12-08-00	00-00-01-01-08-15-54	41.9	5/5
5	0	3	56/84	03-10-21-15-02-03-01	03-00-01-44-24-11-01	93.8	3/5
6	0	3	63/84	13-16-13-08-11-05-01	02-09-05-05-16-24-23	33.4	5/5
7	1	1	55/84	03-09-09-13-16-05-00	00-02-07-04-10-13-48	40.7	5/5
8	1	1	59/84	06-17-14-06-09-06-02	00-01-02-02-10-20-49	36.5	5/5
9	1	2	48/84	15-13-08-06-10-07-03	04-06-05-01-08-13-47	47.8	4/5
10	1	2	57/84	14-06-11-02-14-06-03	00-00-06-03-11-13-51	48.8	5/5
11	1	3	63/84	22-17-16-13-10-06-00	18-25-18-09-10-04-00	54.2	5/5
12	1	3	57/84	10-11-13-16-05-07-00	00-01-01-24-07-09-42	47.8	-
13	2	1	53/84	05-14-18-07-06-07-01	27-12-03-05-06-05-26	61.5	3/5
14	2	1	50/84	00-04-09-08-14-10-06	00-00-00-02-03-12-67	94.5	5/5
15	2	2	56/84	12-08-07-08-08-08-08	02-03-08-04-11-12-44	56.0	5/5
16	2	2	48/84	06-07-09-05-07-09-05	04-02-06-06-07-11-48	79.2	5/5
17	2	3	52/84	05-06-03-06-10-10-11	00-00-02-08-10-19-45	52.9	4/5
18	2	3	51/84	17-07-14-01-05-05-03	00-00-03-09-02-06-64	56.6	5/5
19	3	1	52/84	13-11-07-08-12-05-00	00-00-01-03-14-10-56	68.3	5/5
20	3	1	58/84	04-07-07-05-12-11-11	18-05-03-10-09-04-35	53.9	4/5
21	3	2	60/84	06-06-09-15-12-12-04	00-01-06-06-07-05-59	56.7	5/5
22	3	2	49/84	04-08-12-10-06-04-07	01-07-06-07-09-13-41	67.2	5/5
23	3	3	53/84	10-11-16-05-07-05-02	04-03-04-10-21-15-27	73.0	5/5
24	3	3	61/84	09-17-14-14-04-02-01	08-06-07-05-07-12-39	39.4	4/5
25	4	1	61/84	14-21-10-06-07-01-01	17-12-10-12-14-08-11	36.6	5/5
26	4	1	58/84	02-08-03-02-14-14-15	01-06-06-02-13-25-31	49.0	5/5
27	4	2	58/84	14-16-10-08-05-03-01	01-01-02-05-08-17-49	51.0	4/5
28	4	2	58/84	06-06-16-10-10-06-04	04-04-04-06-09-10-47	46.1	5/5
29	4	3	56/84	07-21-13-05-09-02-03	06-04-05-00-11-07-51	50.0	5/5
30	4	3	42/84	03-11-02-07-09-08-00	21-10-10-02-08-15-18	50.8	5/5
31	5	1	38/84	04-09-04-07-08-04-03	00-01-01-04-04-04-70	125.6	4/5
32	5	1	50/84	03-07-06-11-12-10-02	00-02-10-08-16-26-22	53.2	5/5
33	5	2	57/84	15-10-11-10-10-03-00	05-03-02-18-08-06-42	55.5	5/5
34	5	2	56/84	09-18-14-11-05-01-00	01-03-07-05-07-10-51	51.0	5/5
35	5	3	44/84	07-07-08-11-05-02-00	00-00-02-09-41-15-17	73.2	4/5
36	5	3	56/84	11-10-08-11-09-09-08	03-10-08-07-13-22-21	60.3	5/5
37	6	1	58/84	08-14-12-11-08-04-00	00-07-13-06-28-21-09	53.8	4/5
38	6	1	51/84	16-19-15-04-02-00-00	01-05-07-09-12-20-30	72.2	4/5
39	6	2	53/84	10-06-10-05-14-07-03	00-00-02-05-08-15-54	42.0	-
40	6	2	46/84	01-06-06-03-15-11-04	04-06-10-02-14-06-42	54.6	5/5
41	6	3	51/84	14-09-08-07-08-07-01	05-06-01-03-06-02-61	35.5	-
42	6	3	48/84	07-12-07-10-09-04-00	02-05-09-03-17-13-35	40.0	5/5
43	7	1	38/71	14-14-07-07-03-00-00	03-05-03-04-04-15-37	39.7	4/5
44	7	1	40/71	01-07-09-05-07-05-05	03-05-10-06-15-18-14	32.2	5/5
45	7	2	47/71	06-08-09-13-13-02-00	01-07-08-05-16-32-02	31.3	5/5
46	7	2	55/71	12-10-13-08-10-06-00	03-04-10-01-23-10-20	60.3	4/5
47	7	3	45/71	05-20-11-04-05-00-00	00-06-05-02-04-11-43	44.9	5/5
48	7	3	46/71	08-13-14-05-04-02-00	03-03-05-02-05-19-34	33.8	5/5

7.B SEGMENTATION NETWORK DETAILS

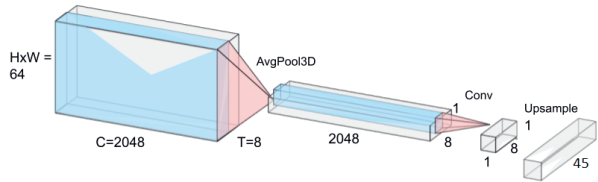
Figure 7.10 presents the architecture of the segmentation heads used. For all models, we apply pooling and convolution operations over the spatial and channel dimensions and up-sample the time dimension to the length of the target segmentation mask (45). Output masks are averaged for multimodal methods.



(a) Time series ResNet head (acceleration modality).



(b) Audio ResNet head.



(c) Video ResNet (slow model) head.

Figure 7.10: Segmentation heads for acceleration, audio, and video models. The first block represents the feature map before the head of the ResNet model, for each modality method. Subsequent operations pool and convolve over the spatial and channel dimensions, and up-sample the time dimension.

8

DISCUSSION AND FUTURE WORK

In this final chapter, we discuss the thesis contributions with respect to limitations and challenges towards the machine assessment of social experience in mingling settings, as discussed in Chapter 1.

8.1 SUMMARY OF CONTRIBUTIONS AND FINDINGS

We start by reviewing the work per chapter. In this thesis, we presented the following contributions:

Exploration of the link between body acceleration and attraction (Chapter 2) presents a study of attraction in the dyadic speed date setting. This work focused on the link between accelerometer signals, capturing overall body movement and attraction. Its main contribution is evidence that, in the dyadic speed dating setting, it is possible to detect constructs like attraction (self-reported) from raw body movement signals obtained from a chest-worn accelerometer. We used hand-crafted features, inspired by previous work, to capture individual and pairwise body movement information. The pairwise features used in our study were designed to capture synchrony, mimicry, and convergence information. Our work therefore also tested and ultimately supported the link between these concepts (as captured by our features) and attraction, a link supported by previous work. Machine learning experiments revealed that both individual and pairwise features were similarly predictive of attraction ratings, but joint features delivered better performance. Statistical analysis suggested that the overall increase or decrease in an individual's body movement throughout an interaction is a potential indicator of multiple types of attraction. Interactions in which the female sought friendship or the male sought romantic or sexual goals had a more characteristic signature in individual body movement. An ablation study comparing the different feature sets revealed that convergence features, designed to measure the degree to which the body movement of both subjects grew similar throughout the interaction were more predictive than features designed to capture mimicry and synchrony. Our results, obtained in relatively constrained pairwise

interaction provide support for the idea that assessing social experience constructs directly from (limited) body movement information (Section 1.2) is a valid approach.

Speaking status detection through feature filtering (Chapter 3) is a contribution towards the detection of speaking status in crowded scenes, by minimizing the adverse effect of cross-contamination in video (the fact that bodies often overlap in a video of a crowd) through two avenues: the use of accelerometer as an alternative modality; and the use of a filtering approach using body pose estimation to exclude contaminated features.

ConfLab: data collection for analysis of mingling settings in the wild (Chapter 4) presents a new data collection concept, dataset, and benchmark for machine analysis of mingling settings. The ConfLab dataset was collected in the wild during a machine learning conference event. This dataset is unique among mingling datasets for its conference setting, and for having been annotated for body joints.

In addition to the dataset, ConfLab introduced methodological improvements to the data collection and annotation processes for mingling settings via key design decisions and technological innovations: a) camera setup and synchronization, b) higher-fidelity and smaller-form wearable badges, c) continuous annotation of poses and actions. All details of our techniques were also made freely available to aid future data collection efforts. Our sensor setup improved upon the data fidelity of prior in-the-wild datasets while retaining privacy sensitivity: 8 videos (1920 × 1080, 60 fps) from a non-invasive overhead view, and custom-made wearable sensors with on-board recording of body motion (full 9-axis IMU), privacy-preserving low-frequency audio (1250 Hz), and Bluetooth-based proximity. Body joint annotations opened the door to studying tasks previously unexplored in the in-the-wild mingling datasets. The dataset baselines showcased three such tasks: body joint detection from overhead camera views, pose-based no-audio speaker detection, and F-formation detection from orientation information. These tasks bridge the gap between the existing fields of pose estimation and pose-based action recognition; and mingling settings. Until now, it has been impossible to evaluate methods from these fields on an in-the-wild social interaction dataset. Other tasks not introduced in the paper such as joint pose-video action recognition are also possible thanks to the availability of poses.

REWIND dataset: a mingling dataset with high-quality individual audio (Chapter 5)

Being the first mingling dataset providing high-quality individual audio recordings from 33 data subjects, REWIND fills a gaping hole in the study of mingling settings. First, the availability of audio makes it possible to annotate vocal or multi-modal behavior (eg. speech, laughter, back-channeling) with access to speech / vocal production. In addition to the annotation of the dataset itself, this provides a way to validate the methods used by previous works, which had to default to the use of video alone for annotation of primarily vocal phenomena such as speech. Beyond annotation, the availability of audio enables the study of such behaviors across modalities, including the development of audiovisual detection models. Finally, the availability of a diverse set of body movement modalities: video, poses, and acceleration make the dataset an attractive in-the-wild benchmark for body-movement-based speaking

status detection. Given the task's difficulty (as shown by our benchmarks), fully exploiting these three modalities is attractive for future work.

Covfee: a web software framework for continuous annotation (Chapter 6) presents the Covfee framework, a web software framework for continuous annotation that both a) implements novel continuous annotation techniques for keypoints and actions, used in the crowd-sourced annotation of Conflab, and b) lowers the bar for experimentation with new kinds of annotation techniques and interfaces. This addresses some of the key challenges faced in the annotation, and the study of annotation of human behavior (Section 1.4.3). The time benefits of the keypoint annotation technique are validated in this work. We measured a three-fold decrease in annotation time with no loss in inter-rater agreement. We additionally present the design requirements, choices, workflow, and features of the Covfee framework that enable both the straightforward use of currently implemented annotation techniques and the implementation of new techniques. These include classes and interfaces for data storage, integration with crowd-sourcing workflows, support for automated annotator qualification tests, and annotation tracking and monitoring features. The Covfee framework is freely available online under a free software license. We expect that providing a platform for rapid prototyping of continuous annotation techniques in addition to showing the feasibility of such techniques will further motivate researchers to follow up on this line of work, to address the shortcomings of existing techniques (Section 1.4.3)

Exploration of differences in the annotation of laughter across modalities (Chapter 7)

focused on the effect of the availability of different modalities during annotation. This is done in the context of laughter. Although laughter is well-recognized as a multimodal phenomenon, it is unclear how annotation of laughter differs when done from modalities like video, without access to audio. This is particularly relevant for in-the-wild mingling dataset, where audio recordings are often unavailable. In this paper, we take a first step in this direction by asking if and how well laughter can be annotated when only audio, only video (containing full body movement information) or audiovisual modalities are available to annotators. We ask whether annotations of laughter are congruent across modalities, and compare the effect that labeling modality has on machine learning model performance. We compare annotations and models for laughter detection, intensity estimation, and segmentation, three tasks common in previous studies of laughter. Our analysis is in the context of a challenging in-the-wild conversational dataset with a variety of camera angles, noise conditions, and voices. Our statistical analysis of more than 4000 annotations acquired from 48 annotators revealed that laughter could be annotated from video with high precision. Recall, lower on average than for the audio and audiovisual conditions, tended to increase with the intensity of the laughter samples. Inter-annotator agreement revealed evidence for a discrepancy in the perception of laughter and its intensity between modalities. Our machine learning experiments compared the performance of unimodal (audio-based, video-based, and acceleration-based) and multi-modal models for different combinations of input modalities, training label modality, and testing label modality, for more than 120 model evaluations in total.

These results revealed that training labels acquired in the audio and audiovisual conditions resulted in the best model performances (mirroring the agreement scores). However, models with video and acceleration inputs were consistent regardless of training label modality, suggesting that it is appropriate to train state-of-the-art models for laughter detection from body movements using video-acquired labels.

8.2 DISCUSSION OF IMPLICATIONS

In Section 1.4, we presented challenges in three stages of social signal processing studies: data collection, annotation, and modeling/analysis. Here we discuss the implications of the works presented in this thesis towards each stage. Note that this separation was done to provide more structure to the discussion and is treated loosely. These stages are not independent and we break boundaries in the discussions where necessary.

8.2.1 DATA COLLECTION

The scaling of data collection efforts faces challenges that seem hard to circumvent, especially related to privacy restrictions stemming from the extra-personal nature of social information (Section 1.4.2). Consumer wearable devices are being used for the large-scale collection of personal health and behavior information [382]. Studies have recorded physiological signals for thousands of subjects throughout weeks or months [383]. Similar to our challenge of assessing social experience, wearable signals have been used to automatically assess individual experience constructs, such as well-being [384] and engagement with an activity [385], recorded over days or weeks.

However, such a scale is currently out of reach in the study of social signals, and particularly in the mingling setting due to the privacy challenges involved (Section 1.4.2). Motivated researchers will certainly be able to keep collecting datasets of similar scale (as current datasets) using existing practices, and supplying the field with valuable data for scientific study. However, thinking about the goal of social signal processing of ultimately aiding human interaction through interventions begs the question: *will socially intelligent systems ever be widely deployed in mingling settings under the current privacy (law) landscape?* This seems an unlikely possibility for systems based on current data collection techniques, due to the issue of *unanimous consent*. Grounds for data processing other than consent are hard to argue for non-essential data processing for research purposes. A reformulation of data collection practices around privacy is likely necessary for systems to be widely deployable. Such proposals have already been made in the robotics field [386], which faces similar challenges.

Such a future may involve privacy-preserving modalities that can collect rich social information from a subject (the consent-giver) while avoiding the capture of personal data from non-consenting others. Some of the contributions presented in this thesis, such as the improvements to wearable devices developed for Conflab may contribute towards such a future. In particular, the wearable badges presented with Conflab significantly improve the recording of individual signals. The IMU signals recorded by the wearable capture only the wearer, and have privacy advantages when recording a scene without consent from all subjects, or for longitudinal recordings (Section 1.4.2). Acceleration signals, however, are low-dimensional when compared to video and the fact that they capture only the wearer

means that they carry limited social information.

Other modalities captured in Conflab such as Bluetooth-based proximity are a first step towards capturing information about social connections while maintaining privacy. A possibly harder question to answer is whether it is possible to collect sufficient social information for meaningful analyses without collecting personal data at all. Such questions have also been asked in the field of robotics [387, 388]. Answering this question precisely, however, will require that researchers take a stance (directly or implicitly) about what constitutes personal data exactly (ie. what is an acceptable threshold for identifiability). The law does not currently give a detailed answer to this question from the point of view of AI, as the means for identifying a subject are largely unspecified. Some scholars have argued that virtually any data can be considered personal since it can be associated with a person given enough processing power and auxiliary data. Law experts have argued that this aspect of GDPR is being challenged by AI and Big Data [387]. Other areas of GDPR may be challenged too. A second critical point regards the definition of data processing. Currently, data processing is understood, under GDPR, to include virtually any form of data manipulation. Therefore data processing for immediate anonymization, even without persistent storage of the personal data, is subject to consent. This point could be challenged in the future by the low risk of immediate anonymization and the many possibilities it creates. A change on this point alone could allow for the recording of video (for research purposes) without consent if its immediate anonymization is possible via techniques such as pose detection, facial blurring, or inpainting, already the subject of extensive research [389].

Meanwhile, the more traditional audio and video modalities face large privacy and scalability challenges. Recording audio in a mingling setting is hard to achieve without cross-contamination, where it is challenging to limit the recording to the wearer of the microphone, who provided consent. Future work is necessary to address this challenge. Hardware such as beam-forming microphones or throat microphones could provide a suitable solution in the future.

The video modality is perhaps more ubiquitous than audio in mingling settings and is almost universally the modality of choice for annotation of body actions (eg. laughter, nodding, gestures). There is an open challenge in finding more privacy-sensitive alternatives suitable for annotation (Section 1.4.2). In this thesis (Chapter 4) it was shown that acceleration can be used to detect speaking (and speaking-related gestures) in a mingling setting with AUCs of around 0.8, which, depending on the application, could be sufficient for many downstream tasks. However, less common signals such as laughter or back-channeling are harder to separate from the noise in an acceleration signal. Differentiating gestures is also hard to accomplish without multiple devices. Alternatives to video could involve one or a combination of carefully placed video cameras (face recordings are considered personal, but the same may not be true of the rest of the body) and higher-quality wearable sensors in larger numbers. It is possible that wrist accelerometers in addition to chest-worn ones would significantly improve laughter detection performance.

Beyond the mingling setting, online crowdsourcing made more popular with the Covid-19 pandemic, can offer a more efficient way to collect and annotate data of social interactions at a scale hard to reach offline. The benefits of crowd-sourced data collection come from the lower need for instrumentation, access to large subject pools, and straightforward online

elicitation of consent. While the video call setting is fundamentally different from offline interaction, the practical benefits of researching this setting at a larger scale, together with its growing importance in our lives, could propel it forward as a primary setting of interest. Here, social signal processing may, in the future, be able to achieve its goals of aiding social experience through interventions and discovering relevant social facts about human (sub-)populations. New software tools are needed to aid the process, with attention to user experience, correct ethical and data management procedures, coverage of a wide range of data collection, self-reporting and annotation needs, and attention to code quality and re-usability. The Covfee framework presented in this thesis aims to be an example in these areas. In particular, Covfee provides many building blocks to support online data collection (eg. via webcam recordings). Expanding the platform in that direction could enable efficient, reproducible, and scalable human behavior data collection methodologies to be deployed on crowd-sourcing markets. No standard tools exist to fill this gap.

8.2.2 ANNOTATION

In this section, we refer to the implications of this thesis (particularly Chapters 6 and 7) in the data annotation stage (Section 1.4.3).

As with data collection, social signal processing faces the challenge of supporting the annotation of larger datasets with more diversity of subjects, demographics and settings. Contributions towards improving the time efficiency of annotations are particularly important in this goal (Section 1.4.3). Continuous annotation has been presented in this work as a promising technique for improving annotation time efficiency, possibly at the expense of some time precision. While we showed some early promising results, many research questions about the quality and time efficiency of continuous behavior annotation remain unanswered. Only in the case of action annotation, the variety in the rate of occurrence and characteristics of social actions (eg. speech, laughter, back-channeling) nuances research in this topic. Questions of interest include: can continuous action annotation offer significant time improvements over previous approaches? how are annotation time and quality affected by action characteristics, such as rate of occurrence? How to optimally address annotation delay for different types of actions? Can action classification be done at the same time? Is it feasible for annotators to localize different actions at the same time (eg. speech and laughter)? If so, how many? How would all these factors affect annotator fatigue? Is it possible to improve the process by using different input devices such as gamepads or specialized devices?

Similarly, applying continuous annotation to body joint annotation creates its own questions, starting with: is it necessary to extensively annotate mingling videos for body joints? An automatic solution to the problem would be ideal. While current pose detectors struggle with the side-elevated and top-down views commonly used in mingling settings, they work well with frontal shots of subjects. In contrast with the action annotation challenge, where action recognition methods struggle to detect social actions in mingling settings regardless of the viewpoint; pose detectors achieve near-perfect pose action recognition in frontal shots. Model improvements and the use of larger side-elevated and top-down image datasets may be enough to solve the problem to a sufficient standard on top-down views. This is still a large undertaking, prompting the current attractiveness of manual annotation. Beyond this basic question, there are others such as: how high an

annotation frequency is necessary for the analysis of social phenomena from body poses? Which body joints are more predictive of different social constructs?

Successfully answering many such research questions involves considering the needs of the machine learning system and, crucially, of the annotator. Therefore, questions such as these should ideally be answered holistically, combining methodologies from the human-computer interaction and machine-learning communities.

Beyond continuous annotation, techniques such as semi-automatic, model-assisted annotation, model bootstrapping, or active learning are other possibilities for enhancing the annotation process. Particularly, a great need exists to bring about the engineering solutions necessary to effectively research these subjects. Tooling needs are often research-question specific, and existing tools fail to provide the flexibility necessary to apply to a wide range of research questions. The Covfee platform aims to be an example in this point by providing a flexible platform instead of a single specific tool. However, more development will be needed to implement annotation techniques on top of platforms such as Covfee and to extend such platforms with features to support algorithmic assistance in annotation, active learning, and similar techniques.

Improving time efficiency is far from the only open challenge in social behavior annotation. The challenge of data capture (ie. modalities, settings, sensor setup) for social signals is also a question about annotation (Section 8.2.1). In Chapter 7 we showed that laughter can be annotated from video alone in a mingling dataset, a setting especially challenging for video-based annotation. These results provided a degree of validation for using video-only annotation in existing and future datasets. They also open the door for work studying further questions about the effect of variables affecting the observability of behavioral cues, such as video quality and occlusion.

Furthermore, the machine learning results in this work could have large implications in labeling for action recognition model training, as they indicate that labeling from video, despite the lower agreement of the resulting annotations, may not affect the performance of models trained on inputs containing body movement information (video and acceleration). This finding challenges the idea that laughter, studied as a primarily vocal phenomenon should always be annotated from audio. This draws attention to the fact that inter-rater agreement measures are designed to do exactly what their name implies: measure agreement between annotations. They may not be a useful guideline for selecting annotation procedures to maximize model performance. Our results beg questions like: *when is it better (for performance) to train on a modality with lower inter-rater agreement? Is it generally optimal to train with labels acquired from a modality closely matching the model's input modality (as we observed), even though their inter-rater agreement might be low?* Beyond laughter, answering these questions would inform the development of models for the detection of multimodal social signals like back-channels or speech-related gestures.

8.2.3 MODELLING AND ANALYSIS

Rather than being independent, the stages of data collection, annotation, and modeling are interrelated and involve complex trade-offs. Decisions at the data collection stage impact the latter stages of annotation, modeling, and analysis. Works not primarily concerned with modeling, such as the datasets and annotation studies presented in the previous sections are very much capable of challenging our assumptions about *what social signals to model*, a

question at least as important as the more commonly-addressed *how to model social signals*; to the extent that they can be considered separate. In this section, we address how the work in this thesis impacts our knowledge about the modeling of social signals, and we address important open challenges faced in the field.

The suggestion presented in Chapter 2 that a phenomenon as subtle as attraction may be manifest in overall body movement is a promising prospect due to attraction being a subtle high-level construct that may be hard to detect even for the humans involved in the interaction. The use of accelerometers also has the privacy advantages discussed in Section 1.4.2 in addition to being an unobtrusive, inexpensive, and low-dimensional modality. More work is necessary to understand the capabilities of accelerometers (and other sensors) in capturing such high-level constructs over a conversation.

Understanding the capabilities of sensor devices is essential since they constitute the measuring devices for social signals. Much in the way that other disciplines calibrate the sensitivity of their devices and calculate error rates for the detection of phenomena of interest, it may be desirable to develop in the direction of obtaining an in-depth understanding of the capabilities of sensors and their associated machine learning models. When research questions revolve around the capacity of sensors to capture social signals or higher-level constructs, rigorously controlled studies could be in order before experiments are performed in the wild. This is because controlling for potential confounding variables like demographics and especially interaction type and conditions could be necessary to understand the capabilities of a sensor. This is particularly true for effects expected to be subtle, such as the effect of acceleration on attraction, where a large set of equal-length dyadic conversations would presumably expose the effect better than in-the-wild unconstrained interactions of varying lengths and group sizes. Physical conditions likely play an important role too. The fact that participants were seated during the speed dates, for example, imposed a constraint on their movements that could have had an important effect on the study results. Position, attachment, and calibration of the accelerometer devices are also important methodological decisions that should be carefully considered in work of this kind. The resolution of the accelerometer devices may also affect its ability to capture subtle movements, potentially related to breathing patterns. Adding accelerometers on other body parts such as the head or wrists may allow capturing a more complete representation of body movement, at the expense of making synchronization necessary. Such rigorous systematic testing of the capabilities of sensor modalities should not be unique to accelerometers. Addressing this challenge requires interdisciplinary work, likely through collaboration between experts in the sensors themselves and experts in the social sciences.

The results from Chapter 2 contributed to a body of work studying the link between synchrony and variables such as cohesion [136], affect [66, 137], attraction [56] or relationship quality [138]. Synchrony is almost exclusively studied in relation to another variable due to the absence of a specific and accepted coding scheme for synchrony. This has created a situation where most work on synchrony creates its own definition of it. Furthermore, most of these definitions correspond to specific engineered features such as those used in Chapter 2. Of particular interest looking forward could be operational definitions that make use of deep learning methods to capture unobserved constructs. Certain basic definitions of synchrony such as the "predictability of subject A's behavior given subject B's behavior" could lend itself to model-based operational definitions of synchrony, where

a model optimized for predicting behavior across subjects is used to quantify the level of synchrony. This idea immediately raises other questions such as: is it wise to define synchrony in terms of predictability or mutual information between signals? How do we deal with difficult cases, eg. are two people who are not moving exhibiting movement synchrony or not? Is most variance in synchrony not entirely context-dependent rather than interaction-dependent?

Many open questions remain about synchrony, including its relevance in real-life social interactions such as mingling settings. While certain dimensions of synchrony such as mimicry are relevant predictors of higher-level constructs in constrained lab settings, the same degree of predictive power has not been observed in the wild. This may be due to limitations in current data collection procedures, or simply because any structure measurable in social behavior as a result of synchrony is small in comparison with the general unpredictability of social behavior. However, since the predictive power of synchrony in the wild is likely to be low, work towards understanding it should

With the ConfLab and REWIND datasets, we collected sets of modalities that directly address open modeling challenges in our field, particularly centered around the recognition of actions and the direct assessment of social experience. As discussed in section 1.4.3 pose annotations (provided for both datasets) are useful as an input modality that enables more detailed analysis when compared to traditional bounding boxes. However, it should not be assumed that pose annotations are necessary or optimal for our goal given how little we understand about the optimal ways to detect actions in social settings. While the merits of acceleration-based detectors have been established in previous work and supported in this thesis, social action detection from video and pose faces bigger challenges. The high dimensionality of video in particular means that datasets used in social signal processing are likely too small and low-variance to adequately utilize the power of modern action recognition methods. Meanwhile, poses are hard to annotate with high precision and this may affect the performance of pose-based models.

This points once again stresses the urgency of stepping up data collection efforts to further progress in social signal processing, particularly in niche sub-fields such as the analysis of mingling settings. This does not mean, however, that progress in modeling does not hold relevance. It does mean, however, that the field cannot presently fully take advantage of many of the breakthroughs brought about by large, data-hungry, deep learning action recognition methods. Such methods are seldom evaluated in social interaction datasets partly due to their smaller size and lower variance. It is unclear which architectures and best practices from the action recognition community transfer to social action recognition, and entirely new ideas are likely necessary to address the specific challenges of social settings. In Chapter 3, for example, we showed that it is possible to improve the performance while reducing the input size to a video-based speaking status detection method via leveraging body poses. Although this work has the limitation of not using state of the art video action recognition models, it suggests that filtering the input to the recognition method through the use of poses may be effective in mingling settings. This has implications in the design of such methods, especially for the side-elevated view, since pose tracks can be obtained automatically and cross-contamination and occlusion are prevalent. More work is warranted to understand whether equivalent ideas can be applied successfully to more recent action recognition methods.

The exploration of the transfer of knowledge across social settings (with eg. zero-shot, one-shot, few-shot learning) may be particularly promising in social action recognition in mingling settings. For example, large movie and meeting datasets contain a wealth of social information that models could potentially learn to transfer to naturalistic mingling settings. Training and even extensively testing large action recognition models, however, currently requires significant computational power and is not a possibility for every research center. Ongoing efforts to develop more data-efficient methods are therefore also critical to social signal processing, particularly for the analysis of mingling settings.

BIBLIOGRAPHY

BIBLIOGRAFIE

- [1] Isabella Poggi and Francesca D'Errico. Social Signals: A Psychological Perspective. In *Computer Analysis of Human Behavior*, pages 185–225. Springer London, January 2011. ISBN 978-0-85729-994-9. doi: 10.1007/978-0-85729-994-9_8. URL https://doi.org/10.1007/978-0-85729-994-9_8.
- [2] Judee K. Burgoon, Norah E. Dunbar, and Howard Giles. Interaction coordination and adaptation. *Social Signal Processing*, pages 78–96, 2017. doi: 10.1017/9781316676202.008. ISBN: 9781316676202.
- [3] Robert R. Provine and Yvonne L. Yong. Laughter: A Stereotyped Human Vocalization. *Ethology*, 89(2):115–124, 1991. ISSN 14390310. doi: 10.1111/j.1439-0310.1991.tb00298.x. ISBN: 0179-1613.
- [4] Donald E. Mowrer, Leonard L. LaPointe, and James Case. Analysis of five acoustic correlates of laughter. *Journal of Nonverbal Behavior*, 11(3):191–199, 1987. ISSN 01915886. doi: 10.1007/BF00990237.
- [5] William Curran, Gary J. McKeown, Magdalena Rychlowska, Elisabeth André, Johannes Wagner, and Florian Lingenfelser. Social context disambiguates the interpretation of laughter. *Frontiers in Psychology*, 8:1–12, 2018. ISSN 16641078. doi: 10.3389/fpsyg.2017.02342.
- [6] Julia Vettin and Dietmar Todt. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115, 2004. ISSN 01915886. doi: 10.1023/B:JONB.0000023654.73558.72.
- [7] Ye Tian, Chiara Mazzocconi, and Jonathan Ginzburg. When do we laugh? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, volume Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 360–369. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3645. URL <https://aclanthology.org/W16-3645>.
- [8] Adrienne Wood and Paula Niedenthal. Developing a social functional account of laughter. *Social and Personality Psychology Compass*, 12(4), 2018. ISSN 17519004. doi: 10.1111/spc3.12383.
- [9] Chiara Mazzocconi, Gulun Jin, Vladislav Maraev, Christine Howes, Jonathan Ginzburg, and Sophie Scott. Laughables and laughter perception : Preliminary investigations. In *CLASP Papers in Computational Linguistics, Vol. 2. Dialogue and Perception - Extended papers from DaP2018 / edited by Christine Howes, Simon Dobnik and Ellen Breitholtz*, pages 72–81, Gothenburg, 2020.

- [10] Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. ISSN 1949-3045. doi: 10.1109/TAFFC.2020.2994533.
- [11] Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Tian Ye. Understanding Laughter. *20th Amsterdam Colloquium*, page 11, 2015.
- [12] ELizabeth Holt. Conversation Analysis and Laughter. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd, 2012. ISBN 978-1-4051-9843-1. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0207>. ISBN: 9781405198431.
- [13] Francesca Bonin, Nick Campbell, and Carl Vogel. Time for laughter. *Knowledge-Based Systems*, 71:15–24, 2014. ISSN 09507051. doi: 10.1016/j.knosys.2014.04.031. URL <http://dx.doi.org/10.1016/j.knosys.2014.04.031>. Publisher: Elsevier B.V.
- [14] W. Tecumseh Fitch. The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, 4(7):258–267, July 2000. ISSN 1364-6613. doi: 10.1016/S1364-6613(00)01494-7. URL <https://www.sciencedirect.com/science/article/pii/S1364661300014947>.
- [15] Matthew Gervais and David Sloan Wilson. The evolution and functions of laughter and humor: a synthetic approach. *Q Rev Biol.*, 80(4):395–430, 2005.
- [16] Marina Davila Ross, Michael J Owren, and Elke Zimmermann. Reconstructing the Evolution of Laughter in Great Apes and Humans. *Current Biology*, 19(13): 1106–1111, July 2009. ISSN 0960-9822. doi: 10.1016/j.cub.2009.05.028. URL <https://www.sciencedirect.com/science/article/pii/S0960982209011294>.
- [17] Barbara Wild, Frank A. Rodden, Wolfgang Grodd, and Willibald Ruch. Neural correlates of laughter and humour. *Brain*, 126(10):2121–2138, October 2003. ISSN 0006-8950. doi: 10.1093/brain/awg226. URL <https://doi.org/10.1093/brain/awg226>.
- [18] Jaak Panksepp. The Riddle of Laughter: Neural and Psychoevolutionary Underpinnings of Joy. *Current Directions in Psychological Science*, 9(6):183–186, December 2000. ISSN 0963-7214. doi: 10.1111/1467-8721.00090. URL <https://doi.org/10.1111/1467-8721.00090>. Publisher: SAGE Publications Inc.
- [19] Michael Miller and William F. Fry. The effect of mirthful laughter on the human cardiovascular system. *Medical Hypotheses*, 73(5):636–639, November 2009. ISSN 03069877. doi: 10.1016/j.mehy.2009.02.044.
- [20] Richard E. Heyman, Michael F. Lorber, J. Mark Eddy, and Tessa V. West. Behavioral Observation and Coding. In Charles M. Judd and Harry T. Reis, editors, *Handbook of Research Methods in Social and Personality Psychology*, pages 345–372. Cambridge University Press, Cambridge, 2 edition, 2014. ISBN 978-1-107-01177-9. doi: 10.1017/CBO9780511996481.018. URL <https://www.cambridge.org/core/books/handbook-of-research-methods-in-social-and-personality-psychology/behavioral-observation-and-coding/2F7375E96411E8149612C9C68315F8B3>.

- [21] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the Gap Between Social Animal and\nUnsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing (to appear)*, 3(1): 1–20, 2011. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.27.1949-3045/12/.
- [22] Stéphane Dupont, Hüseyin Çakmak, Will Curran, Thierry Dutoit, Jennifer Hofmann, Gary McKeown, Olivier Pietquin, Tracey Platt, Willibald Ruch, and Jérôme Urbain. Laughter Research: A Review of the ILHAIRE Project. In *Toward Robotic Socially Believable Behaving Systems*, volume 1. Springer, 2016.
- [23] Lee J. Corrigan, Christopher Peters, Dennis Küster, and Ginevra Castellano. Engagement Perception and Generation for Social Robots and Virtual Agents. In Anna Esposito and Lakhmi C. Jain, editors, *Toward Robotic Socially Believable Behaving Systems - Volume I: Modeling Emotions*, Intelligent Systems Reference Library, pages 29–51. Springer International Publishing, Cham, 2016. ISBN 978-3-319-31056-5. doi: 10.1007/978-3-319-31056-5_4. URL https://doi.org/10.1007/978-3-319-31056-5_4.
- [24] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI*, 7(August): 1–21, 2020. doi: 10.3389/frobt.2020.00092.
- [25] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. Improving the Movement Synchrony Estimation with Action Quality Assessment in Children Play Therapy. In *Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21*, pages 397–406, New York, NY, USA, October 2021. Association for Computing Machinery. ISBN 978-1-4503-8481-0. doi: 10.1145/3462244.3479891. URL <http://doi.org/10.1145/3462244.3479891>.
- [26] Lars Steinert, Felix Putze, Dennis Küster, and Tanja Schultz. Towards Engagement Recognition of People with Dementia in Care Settings. In *Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20*, pages 558–565, New York, NY, USA, October 2020. Association for Computing Machinery. ISBN 978-1-4503-7581-8. doi: 10.1145/3382507.3418856. URL <http://doi.org/10.1145/3382507.3418856>.
- [27] Hyukjae Jang, Sungwon P. Choe, Simon N.B. Gunkel, Seungwoo Kang, and Junehwa Song. A System to Analyze Group Socializing Behaviors in Social Parties. *IEEE Transactions on Human-Machine Systems*, 47(6):801–813, 2017. ISSN 21682291. doi: 10.1109/THMS.2016.2634918.
- [28] Radosław Niewiadomski, Jennifer Hofmann, and Jérôme Urbain. Laugh-aware virtual agent and its impact on user amusement. ... *Agents and Multi-Agent ...*, pages 37–39, 2013. URL <http://dl.acm.org/citation.cfm?id=2485018>. ISBN: 978-1-4503-1993-5.
- [29] Willibald F. Ruch, Tracey Platt, Jennifer Hofmann, Radoslaw Niewiadomski, Jerome Urbain, Maurizio Mancini, and Stéphane Dupont. Gelotophobia and the Challenges of Implementing Laughter into Virtual Agents Interactions. *Frontiers in Human*

- Neuroscience*, 8(November):1–12, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00928. URL <http://journal.frontiersin.org/article/10.3389/fnhum.2014.00928/abstract>.
- [30] Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi, Nigel G. Ward, and Tatsuya Kawahara. Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents. *International Workshop on Spoken Dialog Systems*, pages 1–12, 2016. ISSN 13468030. doi: 10.1527/tjsai.C-G31.
- [31] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amier, and Dirk Heylen. The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4, 2008. doi: 10.1021/cr970463w. URL <http://doc.utwente.nl/65319/>. ISBN: 2-9517408-4-0.
- [32] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2011.20. ISBN: 1949-3045.
- [33] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social Attention: Modeling Attention in Human Crowds. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 4601–4607, 2018. ISSN 10504729. doi: 10.1109/ICRA.2018.8460504. arXiv: 1710.04689 ISBN: 9781538630815.
- [34] Radoslaw Niewiadomski, Yu Ding, Maurizio Mancini, Catherine Pelachaud, Gualtiero Volpe, and Antonio Camurri. Perception of intensity incongruence in synthesized multimodal expressions of laughter. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 684–690, 2015. doi: 10.1109/ACII.2015.7344643. Publisher: IEEE ISBN: 9781479999538.
- [35] Andrew M. Nuxoll and John E. Laird. Enhancing intelligent agents with episodic memory. *Cognitive Systems Research*, 17-18:34–48, July 2012. ISSN 13890417. doi: 10.1016/j.cogsys.2011.10.002. URL <http://linkinghub.elsevier.com/retrieve/pii/S1389041711000428><https://linkinghub.elsevier.com/retrieve/pii/S1389041711000428>. ISBN: 1389-0417.
- [36] Diane S Berry and Jane Sherman Hansen. Personality, Nonverbal Behavior, and Interaction Quality in Female Dyads. *Personality and Social Psychology Bulletin*, 26(3):278–292, 1996.
- [37] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Che-touani. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access*, 5:705–721, 2017. ISSN 21693536. doi: 10.1109/ACCESS.2016.2614525. Publisher: IEEE.
- [38] Rod A. Martin and Nicholas A. Kuiper. Daily occurrence of laughter: Relationships with age, gender, and type a personality. *Humor*, 12(4):355–384, 1999. ISSN 09331719. doi: 10.1515/humr.1999.12.4.355.

- [39] Johan C. Karremans and Thijs Verwijmeren. Mimicking attractive opposite-sex others: The role of romantic relationship status. *Personality and Social Psychology Bulletin*, 34(7):939–950, 2008. ISSN 01461672. doi: 10.1177/0146167208316693. ISBN: 0146-1672.
- [40] Sheida White. Backchannels across Cultures : A Study of Americans and Japanese. *Language in Society*, 18(1):59–76, 1989. URL <https://doi.org/10.1017/S0047404500013270>.
- [41] Simone Pika, Elena Nicoladis, and Paula F. Marentette. A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer? *Bilingualism*, 9(3):319–327, 2006. ISSN 13667289. doi: 10.1017/S1366728906002665.
- [42] Sotaro Kita. Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24(2):145–167, February 2009. ISSN 0169-0965, 1464-0732. doi: 10.1080/01690960802586188. URL <http://www.tandfonline.com/doi/abs/10.1080/01690960802586188>.
- [43] AS Manstead and Agneta H Fischer. Social appraisal. *Appraisal processes in emotion: Theory, methods, research*, pages 221–232, 2001. Publisher: Oxford University Press New York, NY.
- [44] Rafael A. Calvo and Sidney D’Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, January 2010. ISSN 1949-3045. doi: 10.1109/T-AFFC.2010.1. Conference Name: IEEE Transactions on Affective Computing.
- [45] Florian Lingenfelser, Johannes Wagner, Elisabeth André, Gary McKeown, and Will Curran. An event driven fusion approach for enjoyment recognition in real-time. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pages 377–386, 2014. doi: 10.1145/2647868.2654924. ISBN: 9781450330633.
- [46] Siân E. Lindley and Andrew F. Monk. Measuring social behaviour as an indicator of experience. *Behaviour & Information Technology*, 32(10):968–985, October 2013. ISSN 0144-929X. doi: 10.1080/0144929X.2011.582148. URL <https://www.tandfonline-com.tudelft.idm.oclc.org/doi/full/10.1080/0144929X.2011.582148>. Publisher: Taylor & Francis.
- [47] Kevin Doherty and Gavin Doherty. Engagement in HCI: Conception, theory and measurement. *ACM Computing Surveys*, 51(5):1–39, 2019. ISSN 15577341. doi: 10.1145/3234149.
- [48] Nadine Glas and Catherine Pelachaud. Definitions of engagement in human-agent interaction. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 944–949, 2015. doi: 10.1109/ACII.2015.7344688. Publisher: IEEE ISBN: 9781479999538.
- [49] Akilesh Rajavenkatanarayanan, Konstantinos Tsiakas, Ashwin Ramesh Babu, and Fillia Makedon. Monitoring task engagement using facial expressions and body

- postures. *ACM International Conference Proceeding Series*, pages 103–108, 2018. doi: 10.1145/3191801.3191816. ISBN: 9781450354394.
- [50] Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*, pages 440–445, 2019. doi: 10.1145/3340555.3353765. ISBN: 9781450368605.
- [51] Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. Towards the automatic detection of involvement in conversation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6800 LNCS:163–170, 2011. ISSN 03029743. doi: 10.1007/978-3-642-25775-9_16. ISBN: 9783642257742.
- [52] Catharine Oertel, Stefan Scherer, and Nick Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1541–1544, 2011.
- [53] Catharine Oertel and Giampiero Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pages 99–106, 2013. doi: 10.1145/2522848.2522865. ISBN: 9781450321297.
- [54] Ronald Böck, Ingo Siegert, Andreas Wendemuth, and Stefan Glüge. Annotation and classification of changes of involvement in group conversation. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 803–808, 2013. doi: 10.1109/ACII.2013.150. ISBN: 9780769550480.
- [55] Arno Veenstra and Hayley Hung. Do they like me? Using video cues to predict desires during speed-dates. *Proceedings of the IEEE International Conference on Computer Vision*, pages 838–845, 2011. doi: 10.1109/ICCVW.2011.6130339. ISBN: 9781467300629.
- [56] E. Prochazkova, E. Sjak-Shie, F. Behrens, D. Lindh, and M. E. Kret. Physiological synchrony is associated with attraction in a blind date setting. *Nature Human Behaviour*, November 2021. doi: 10.1038/s41562-021-01197-3. URL <https://hdl.handle.net/1887/3248710>.
- [57] Jose David Vargas Quiros, Oyku Kapcak, Hayley Hung, and Laura Cabrera-Quiros. Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. ISSN 1949-3045. doi: 10.1109/TAFFC.2021.3138349. Conference Name: IEEE Transactions on Affective Computing.
- [58] Soumia Dermouche and Catherine Pelachaud. Engagement Modeling in Dyadic Interaction. In *2019 International Conference on Multimodal Interaction, ICMI '19*, pages 440–445, New York, NY, USA, October 2019. Association for Computing Machinery. ISBN 978-1-4503-6860-5. doi: 10.1145/3340555.3353765. URL <http://doi.org/10.1145/3340555.3353765>.

- [59] Jessica L. Lakin and Tanya L. Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14(4):334–339, 2003. ISSN 09567976. doi: 10.1111/1467-9280.14481.
- [60] R. Matthew Montoya, Christine Kershaw, and Julie L. Prosser. A meta-analytic investigation of the relation between interpersonal attraction and enacted behavior. *Psychological Bulletin*, 144(7):673–709, 2018. ISSN 00332909. doi: 10.1037/bul0000148.
- [61] Sally D. Farley. Nonverbal Reactions to an Attractive Stranger: The Role of Mimicry in Communicating Preferred Social Distance. *Journal of Nonverbal Behavior*, 38(2): 195–208, 2014. ISSN 01915886. doi: 10.1007/s10919-014-0174-4.
- [62] Marielle Stel and Roos Vonk. Mimicry in social interaction: Benefits for mimickers, mimicked, and their interaction. *British Journal of Psychology*, 101:311–323, 2010. ISSN 16130073. doi: 10.1348/000712609X465424. ISBN: 0007-1269.
- [63] Tanya L. Chartrand and Amy N. Dalton. Mimicry: Its Ubiquity, Importance, and Functionality. In *Oxford handbook of human action*, pages 458–483. Oxford University Press, 2009. ISBN 0-19-530998-7 (Hardcover); 978-0-19-530998-0 (Hardcover).
- [64] Carlos Cornejo, Zamara Cuadros, Ricardo Morales, and Javiera Paredes. Interpersonal coordination: Methods, achievements, and challenges. *Frontiers in Psychology*, 8 (SEP):1–16, 2017. ISSN 16641078. doi: 10.3389/fpsyg.2017.01685.
- [65] Fabian Ramseyer. Synchronized movement in social interaction. *ACM International Conference Proceeding Series*, pages 1–5, 2013. doi: 10.1145/2557595.2557597. ISBN: 9781450325813.
- [66] Wolfgang Tschacher, Georg M. Rees, and Fabian Ramseyer. Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, 5(NOV):1–13, 2014. ISSN 16641078. doi: 10.3389/fpsyg.2014.01323. ISBN: 1664-1078.
- [67] Ning Yang, Zhelong Wang, and Weijian Hu. Synchrony expression analysis of human-human interaction using wearable sensors. *IEEE International Conference on Control and Automation, ICCA*, 2016-July:611–615, 2016. ISSN 19483457. doi: 10.1109/ICCA.2016.7505345. ISBN: 9781509017386.
- [68] Jane Paulick, Anne Katharina Deisenhofer, Fabian Ramseyer, Wolfgang Tschacher, Kaitlyn Boyle, Julian Rubel, and Wolfgang Lutz. Nonverbal synchrony: A new approach to better understand psychotherapeutic processes and drop-out. *Journal of Psychotherapy Integration*, 28(3):367–384, 2018. ISSN 15733696. doi: 10.1037/int0000099.
- [69] Alexandra Paxton and Rick Dale. Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, 66(11):2092–2102, 2013. ISSN 17470218. doi: 10.1080/17470218.2013.853089.
- [70] Désirée Schoenherr, Jane Paulick, Susanne Worrack, Bernhard M. Strauss, Julian A. Rubel, Brian Schwartz, Anne Katharina Deisenhofer, Wolfgang Lutz, Ulrich Stangier,

- and Uwe Altmann. Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony. *Behavior Research Methods*, 51(1):361–383, 2019. ISSN 15543528. doi: 10.3758/s13428-018-1139-z.
- [71] Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007. ISSN 03701573. doi: 10.1016/j.physrep.2006.11.001.
- [72] Moreno I Coco, Rick Dale, and Noel Nguyen. Cross-recurrence quantification analysis of categorical and continuous time series : an R package. *Frontiers in Psychology*, 5 (June):1–14, 2014. doi: 10.3389/fpsyg.2014.00510.
- [73] Wen-Sheng Chu, Jiabei Zeng, Fernando De La Torre, Jeffrey F Cohn, and Daniel S Messinger. Unsupervised Synchrony Discovery in Human Interaction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3146–3154, 2015.
- [74] Marjolein C. Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlavik, and Hayley Hung. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017-Janua(November): 206–215, 2017. doi: 10.1145/3136755.3136811. ISBN: 9781450355438.
- [75] C. Oertel, J. Gustafson, K.A.F. Mora, and J.-M. Odobez. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 2015. doi: 10.1145/2818346.2820759. ISBN: 9781450339124.
- [76] Catherine Lai, Jean Carletta, and Steve Renals. Modelling Participant Affect in Meetings with Turn-Taking Features. In *Workshop on Affective Social Speech Signals*, 2013. doi: 10.13140/2.1.2624.7042.
- [77] Catherine Lai and Gabriel Murray. Predicting group satisfaction in meeting discussions. *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018*, 2018. doi: 10.1145/3279810.3279840. ISBN: 9781450360722.
- [78] Koji Inoue, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. Annotation and analysis of listener’s engagement based on multi-modal behaviors. In *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction, MA3HMI ’16*, pages 25–32, New York, NY, USA, November 2016. Association for Computing Machinery. ISBN 978-1-4503-4562-0. doi: 10.1145/3011263.3011271. URL <http://doi.org/10.1145/3011263.3011271>.
- [79] Florian Lingensfelder, Johannes Wagner, Jun Deng, Raymond Brueckner, Björn Schuller, and Elisabeth André. Asynchronous and Event-Based Fusion Systems for Affect Recognition on Naturalistic Data in Comparison to Conventional Approaches. *IEEE Transactions on Affective Computing*, 9(4):410–423, October 2018. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2635124. Conference Name: IEEE Transactions on Affective Computing.

- [80] Laura Cabrera-Quiros, David M. J. Tax, and Hayley Hung. Gestures In-The-Wild: Detecting Conversational Hand Gestures in Crowded Scenes Using a Multimodal Fusion of Bags of Video Trajectories and Body Worn Acceleration. *IEEE Transactions on Multimedia*, 22(1):138–147, January 2020. ISSN 1941-0077. doi: 10.1109/TMM.2019.2922122.
- [81] Ekin Gedik and Hayley Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21(4):723–737, 2017. ISSN 16174909. doi: 10.1007/s00779-017-1006-4.
- [82] Kristiina Jokinen, Trung Ngo Trong, and Graham Wilcock. Body Movements and Laughter Recognition: Experiments in First Encounter Dialogues. *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, pages 20–24, 2016. ISSN 00086223. doi: 10.1016/S0008-6223(98)00334-0. URL <http://doi.acm.org/10.1145/3011263.3011264>. ISBN: 8641146719917.
- [83] Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. Decision-level fusion for audio-visual laughter detection. *Lecture Notes in Computer Science*, 5237 LNCS:137–148, 2008. ISSN 03029743. doi: 10.1007/978-3-540-85853-9-13.
- [84] Radoslaw Niewiadomski, Maurizio Mancini, Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. Automated Laughter Detection from Full-Body Movements. *IEEE Transactions on Human-Machine Systems*, 46(1):113–123, 2016. ISSN 21682291. doi: 10.1109/THMS.2015.2480843.
- [85] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*, page 207, 2013. ISSN 0042-4900. doi: 10.1145/2493432.2493513. URL <http://dl.acm.org/citation.cfm?doid=2493432.2493513>. ISBN: 9781450317702.
- [86] Mark A. Schmuckler. What Is Ecological Validity? A Dimensional Analysis. *Infancy*, 2(4):419–436, 2001. ISSN 1532-7078. doi: 10.1207/S15327078IN0204_02. URL http://onlinelibrary.wiley.com/doi/abs/10.1207/S15327078IN0204_02. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/S15327078IN0204_02.
- [87] John F. Kihlstrom. Ecological Validity and “Ecological Validity”. *Perspectives on Psychological Science*, 16(2):466–471, March 2021. ISSN 1745-6916. doi: 10.1177/1745691620966791. URL <https://doi.org/10.1177/1745691620966791>. Publisher: SAGE Publications Inc.
- [88] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2496269.

- [89] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The MatchNMingle Dataset: A Novel Multi-Sensor Resource for the Analysis of Social Interactions and Group Dynamics In-the-Wild During Free-Standing Conversations and Speed Dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2018. ISSN 1949-3045. doi: 10.1109/TAFFC.2018.2848914.
- [90] Hayley Hung and Ben Kröse. Detecting F-formations as dominant sets. *Proceedings of the 13th international conference on multimodal interfaces - ICMI '11*, page 231, 2011. doi: 10.1145/2070481.2070525. URL <http://dl.acm.org/citation.cfm?doid=2070481.2070525>. ISBN: 9781450306416.
- [91] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PLoS ONE*, 10(5):1–26, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0123783. arXiv: 1409.2702 ISBN: 9781450356152.
- [92] Hooman Hedayati and James Kennedy. Comparing F-Formations Between Humans and On-Screen Agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9. Association for Computing Machinery, 2020. ISBN 978-1-4503-6819-3. doi: 10.1145/3334480.3383015. URL <https://doi.org/10.1145/3334480.3383015>. ISBN: 9781450368193.
- [93] Sebastiano Vascon, Eyasu Z. Mequanint, Marco Cristani, Hayley Hung, Marcello Pellillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143: 11–24, 2016. ISSN 1090235X. doi: 10.1016/j.cviu.2015.09.012. Publisher: Elsevier Inc.
- [94] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of F-formations. *Proceedings of the British Machine Vision Conference 2011*, pages 23.1–23.12, 2011. doi: 10.5244/C.25.23. URL <http://www.bmva.org/bmvc/2011/proceedings/paper23/index.html>. ISBN: 1-901725-43-X.
- [95] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ekin Gedik, Ashraful Islam, and Hayley Hung. ConFLab: A Rich Multimodal Multisensor Dataset of Free-Standing Social Interactions in the Wild, July 2022. URL <http://arxiv.org/abs/2205.05177>. arXiv:2205.05177 [cs].
- [96] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. ISSN 02628856. doi: 10.1016/j.imavis.2008.11.007.
- [97] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua(Xxx):1302–1310*, 2017. doi: 10.1109/CVPR.2017.143. arXiv: 1812.08008 ISBN: 9781538604571.
- [98] Hao Shu Fang, Shuqin Xie, Yu Wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. *Proceedings of the IEEE International Conference on Computer Vision*,

- 2017-Octob:2353–2362, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.256. arXiv: 1612.00137 ISBN: 9781538610329.
- [99] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00762. arXiv: 1802.00434 ISBN: 9781538664209.
- [100] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning Individual Styles of Conversational Gesture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL <https://ieeexplore-ieee-org.tudelft.idm.oclc.org/document/8954219>. arXiv: 1906.04160.
- [101] Jiaying Shen, Oren Lederman, Jiannong Cao, Florian Berg, Shaojie Tang, and Alex Sandy Pentland. GINA: Group Gender Identification Using Privacy-Sensitive Audio Data. *Proceedings - IEEE International Conference on Data Mining, ICDM, 2018-Novem:457–466*, 2018. ISSN 15504786. doi: 10.1109/ICDM.2018.00061. Publisher: IEEE ISBN: 9781538691588.
- [102] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. Rhythm: A Unified Measurement Platform for Human Organizations. *IEEE Multimedia*, 25(1): 26–38, 2018. ISSN 1070986X. doi: 10.1109/MMUL.2018.112135958. Publisher: IEEE.
- [103] Jocelynn Cu, Ma Beatrice Luz, Mcanjelo Nocom, and Timothy Jasper Purganan. Affective Laughter Expressions from Body Movements. In *Trends in Artificial Intelligence: PRICAI 2016 Workshops*, 2016.
- [104] L. C. Cabrera Quiros. *Automatic analysis of human social behavior in - the - wild using multimodal streams*. PhD thesis, Delft University of Technology, Delft, 2018. URL <https://repository.tudelft.nl/islandora/object/uuid%3A811ba745-18e1-4dca-8321-249ba000a142>.
- [105] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. URL <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. tex.added-at: 2020-08-20T11:33:21.000+0200 tex.biburl: <https://www.bibsonomy.org/bibtex/243a2175512dc8b9d8855fa7a763cdc3e/zotero> tex.interhash: c7b667cac6031282160a9e94d5a118f8 tex.intrahash: 43a2175512dc8b9d8855fa7a763cdc3e tex.timestamp: 2020-08-20T11:33:21.000+0200.
- [106] Gary McKeown, William Curran, Denise Kane, Rebecca McCahon, Harry J. Griffin, Ciaran McLoughlin, and Nadia Bianchi-Berthouze. Human perception of laughter from context-free whole body motion dynamic stimuli. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 306–311, 2013. ISSN 2156-8103. doi: 10.1109/ACII.2013.57.

- [107] Harry J. Griffin, Min S.H. Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. Laughter Type Recognition from Whole Body Motion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 349–355, September 2013. doi: 10.1109/ACII.2013.64. ISSN: 2156-8111.
- [108] Harry J. Griffin, Min S.Hane Aung, Bernadino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives. *IEEE Transactions on Affective Computing*, 6(2):165–178, 2015. ISSN 19493045. doi: 10.1109/TAFFC.2015.2390627. ISBN: 1949-3045 VO - 6.
- [109] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI Meeting Corpus: A Pre-announcement Machine Learning for Multimodal Interaction. *Machine Learning for Multimodal Interaction SE - Lecture Notes in Computer Science*, 3869:28–39, 2006. doi: doi:10.1007/11677482.3. URL citeulike-article-id:6473361%5Cnhttp://dx.doi.org/10.1007/11677482_3. ISBN: 978-3-540-32549-9.
- [110] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 265–293. PMLR, March 2022. ISSN: 2640-3498.
- [111] Phil Lopes, Georgios N. Yannakakis, and Antonios Liapis. RankTrace: Relative and unbounded affect annotation. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 2018-Janua:158–163, 2018. doi: 10.1109/ACII.2017.8273594. ISBN: 9781538605639.
- [112] B. M. Booth, K. Mundnich, and S. S. Narayanan. A Novel Method for Human Bias Correction of Continuous- Time Annotations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3091–3095, April 2018. doi: 10.1109/ICASSP.2018.8461645. ISSN: 2379-190X.
- [113] Brandon M. Booth and Shrikanth S. Narayanan. Fifty Shades of Green: Towards a Robust Measure of Inter-annotator Agreement for Continuous Signals. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 204–212, Virtual Event Netherlands, October 2020. ACM. ISBN 978-1-4503-7581-8. doi: 10.1145/3382507.3418860. URL <https://dl.acm.org/doi/10.1145/3382507.3418860>.
- [114] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. “Feeltrace”: An instrument for recording perceived emotion in real time. In *ISCA Workshop on Speech & Emotion*, pages 19–24, 2000. doi: citeulike-article-id:3721917.
- [115] Jeffrey M. Girard and Aidan G. Aidan. DARMA: Software for dual axis rating and media annotation. *Behavior Research Methods*, 50(3):902–909, 2018. ISSN

15543528. doi: 10.3758/s13428-017-0915-5. Publisher: Behavior Research Methods
ISBN: 1342801709.
- [116] David Melhart, Antonios Liapis, and Georgios N. Yannakakis. PAGAN : Video Affect Annotation Made Easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019. arXiv: 1907.01008v1.
- [117] Soroosh Mariooryad and Carlos Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, 2015. ISSN 19493045. doi: 10.1109/TAFFC.2014.2334294.
- [118] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. In *AVEC '15: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 41–48, October 2015. ISBN 978-1-4503-3743-4. doi: 10.1145/2808196.2811640.
- [119] Soheil Khorram, Melvin McInnis, and Emily Mower Provost. Jointly Aligning and Predicting Continuous Emotion Annotations. *IEEE Transactions on Affective Computing*, 3045(c):1–16, 2019. ISSN 19493045. doi: 10.1109/TAFFC.2019.2917047.
- [120] Max Planck Institute for Psycholinguistics. ELAN [Computer software]., 2021. URL <https://archive.mpi.nl/tla/elan>.
- [121] Tobias Baur, Ionut Damian, Florian Lingenfeller, Johannes Wagner, and Elisabeth André. Nova: Automated analysis of nonverbal signals in social interactions. In Albert Ali Salah, Hayley Hung, Oya Aran, and Hatice Gunes, editors, *Human Behavior Understanding*, pages 160–171, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02714-2.
- [122] Jason Tipples, Anthony P. Atkinson, and Andrew W. Young. The eyebrow frown: A salient social signal. *Emotion*, 2(3):288–296, 2002. ISSN 1931-1516, 1528-3542. doi: 10.1037/1528-3542.2.3.288. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/1528-3542.2.3.288>.
- [123] María L. Flecha-García. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52(6):542–554, 2010. ISSN 01676393. doi: 10.1016/j.specom.2009.12.003. ISBN: 0167-6393.
- [124] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. PA3D: Pose-action 3D machine for video recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:7914–7923, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00811. ISBN: 9781728132938.
- [125] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Po-Tion: Pose MoTion Representation for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00734. ISBN: 9781538664209.

- [126] Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, and Bjorn Ottersten. ENHANCED TRAJECTORY-BASED ACTION RECOGNITION USING HUMAN POSE. In *IEEE International Conference on Image Processing*, pages 1807–1811, 2017. ISBN 978-1-5090-2175-8.
- [127] Xinsheng Wang, Jihua Zhu, and Odette Scharenborg. Multimodal Fusion of Body Movement Signals for No-audio Speech Detection. In *Working Notes Proceedings of the MediaEval 2020 Workshop*, page 3, 2020.
- [128] Jose Vargas and Hayley Hung. CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection. In *Working Notes Proceedings of the MediaEval 2019 Workshop*, pages 11–13, 2019.
- [129] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864, May 2020. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01273. URL <https://direct.mit.edu/neco/article/32/5/829/95591/A-Survey-on-Deep-Learning-for-Multimodal-Data>.
- [130] UttharaGosa Mangai, Suranjana Samanta, Sukhendu Das, and PinakiRoy Chowdhury. A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. *IETE Technical Review*, 27(4):293, 2010. ISSN 0256-4602. doi: 10.4103/0256-4602.64604. URL <http://tr.ietejournals.org/text.asp?2010/27/4/293/64604>. ISBN: 0256-4602.
- [131] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in A Triadic Interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. URL <http://arxiv.org/abs/1906.04158>. arXiv: 1906.04158.
- [132] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2012.12. Publisher: IEEE ISBN: 2011080053.
- [133] Robert G. Moulder, Steven M. Boker, Fabian Ramseyer, and Wolfgang Tschacher. Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. *Psychological Methods*, 23(4):757–773, 2018. ISSN 1082989X. doi: 10.1037/met0000172.
- [134] F. Behrens, J. A. Snijdwint, R. G. Moulder, E. Prochazkova, E. E. Sjak-Shie, S. M. Boker, and M. E. Kret. Physiological synchrony is associated with cooperative success in real-life interactions. *Scientific Reports*, 10:19609, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76539-8. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7661712/>.
- [135] Scott S. Wiltermuth and Chip Heath. Synchrony and Cooperation. *Psychological Science*, 20(1):1–5, January 2009. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2008.02253.x. URL <https://doi.org/10.1111/j.1467-9280.2008.02253.x>. Publisher: SAGE Publications Inc.

- [136] Joshua Conrad Jackson, Jonathan Jong, David Bilkey, Harvey Whitehouse, Stefanie Zollmann, Craig McNaughton, and Jamin Halberstadt. Synchrony and Physiological Arousal Increase Cohesion and Cooperation in Large Naturalistic Groups. *Scientific Reports*, 8(1):127, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-017-18023-4. URL <http://www.nature.com/articles/s41598-017-18023-4>.
- [137] Bo Xiao, Panayiotis G. Georgiou, Chi Chun Lee, Brian Baucom, and Shrikanth S. Narayanan. Head motion synchrony and its correlation to affectivity in dyadic interactions. *Proceedings - IEEE International Conference on Multimedia and Expo*, pages 2–7, 2013. ISSN 19457871. doi: 10.1109/ICME.2013.6607480. ISBN: 9781479900152.
- [138] F. Ramseyer and W. Tschacher. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 2011. doi: 10.1037/a0023419.
- [139] Tanya L. Chartrand and John A. Bargh. The chameleon effect: The perception-behaviour link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.
- [140] Nicolas Guéguen. Mimicry and seduction: An evaluation in a courtship context. *Social Influence*, 4(4):249–255, 2009. ISSN 15534510. doi: 10.1080/15534510802628173. ISBN: 1090-5138.
- [141] Rick B van Baaren, Rob W Holland, Kerry Kawakami, and Ad van Knippenberg. Research Report Mimicry and Prosocial Behavior. *Psychological Science (Wiley-Blackwell)*, 15(1):71–74, 2004. URL <http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=11862090&site=ehost-live>. ISBN: 09567976.
- [142] Rick B. van Baaren, Rob W. Holland, Bregje Steenaert, and Ad van Knippenberg. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39(4):393–398, 2003. ISSN 00221031. doi: 10.1016/S0022-1031(03)00014-3.
- [143] Shahin Amiriparian, Jing Han, Maximilian Schmitt, Alice Baird, Adria Mallol-Ragolta, Manuel Milling, Maurice Gerczuk, and Björn Schuller. Synchronization in Interpersonal Speech. *Frontiers in Robotics and AI*, 6, 2019. ISSN 2296-9144. doi: 10.3389/frobt.2019.00116. URL <https://www.frontiersin.org/articles/10.3389/frobt.2019.00116/full>. Publisher: Frontiers.
- [144] Joseph Bonito and Joann Keyton. Multilevel measurement models for group collective constructs. *Group Dynamics: Theory, Research, and Practice*, 23:1–21, March 2019. doi: 10.1037/gdn0000096.
- [145] Oyku Kapcak, Jose Vargas-Quiros, and Hayley Hung. Estimating Romantic, Social, and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pages 154–160. Institute of Electrical and Electronics Engineers Inc., September 2019. ISBN 978-1-72813-891-6. doi: 10.1109/ACIIW.2019.8925137.

- [146] Ellen Berscheid and Elaine Hatfield Walster. *Interpersonal Attraction*. Addison-Wesley, Reading, Massachusetts, Menlo Park, California - London - Don Mills, Ontario, 1969.
- [147] R. Matthew Montoya and Robert S. Horton. A Two-Dimensional Model for the Study of Interpersonal Attraction. *Personality and Social Psychology Review*, 18(1):59–86, 2014. ISSN 10888683. doi: 10.1177/1088868313501887.
- [148] M J Rosenberg and C I Hovland. Cognitive, affective, and behavioral components of attitudes. In Milton J Rosenberg, editor, *Attitude organization and change: an analysis of consistency among attitude components*, volume Yale studi, pages 1–14. Yale University Press, New Haven, 1960.
- [149] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009. ISSN 02628856. doi: 10.1016/j.imavis.2009.01.004. URL <http://dx.doi.org/10.1016/j.imavis.2009.01.004>. Publisher: Elsevier B.V. ISBN: 0262-8856.
- [150] Eli J. Finkel, Paul W. Eastwick, and Jacob Matthews. Speed-dating as an invaluable tool for studying romantic attraction: A methodological primer. *Personal Relationships*, 14(1):149–166, 2007. ISSN 13504126. doi: 10.1111/j.1475-6811.2006.00146.x.
- [151] Anmol Madan and Ron Caneel. Voices of attraction, 2004. URL <https://www.media.mit.edu/publications/voices-of-attraction/>.
- [152] Rajesh Ranganath, Daniel Jurafsky, and Daniel Mcfarland. It’s Not You, it’s Me: Detecting Flirting and its Misperception in Speed-Dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 334–342. Association for Computational Linguistics, January 2009. URL <https://aclanthology.org/D09-1035>.
- [153] Raymond Fisman, Sheena S . Iyengar, Emir Kamenica, and Itamar Simonson. Gender Differences in Mate Selection : Evidence from a Speed Dating Experiment. *Oxford University Press*, 121(2):673–697, 2013.
- [154] D. M. Buss and D. P. Schmitt. Sexual strategies theory: an evolutionary perspective on human mating. *Psychological Review*, 100(2):204–232, April 1993. ISSN 0033-295X. doi: 10.1037/0033-295x.100.2.204.
- [155] K. Grammer, M. Honda, A. Juette, and A. Schmitt. Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *Journal of Personality and Social Psychology*, 77(3):487–508, September 1999. ISSN 0022-3514. doi: 10.1037/0022-3514.77.3.487.
- [156] Wim Pouw. Quantifying Gesture-Speech Synchrony. In *6th Gesture and Speech in Interaction Conference*, pages 75–80, Paderborn, 2019. doi: 10.17619/UNIPB/1-815.
- [157] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. ISSN 01676393. doi: 10.1016/j.specom.2013.09.008. ISBN: 0167-6393.

- [158] Craig Foster, Betty Witcher, W. Keith Campbell, and Jeffrey Green. Arousal and attraction: Evidence for automatic and controlled processes. *Journal of Personality and Social Psychology*, 74:86–101, January 1998. doi: 10.1037/0022-3514.74.1.86.
- [159] Gary Lewandowski Jr and Arthur Aron. Distinguishing arousal from novelty and challenge in initial romantic attraction between strangers. *Social Behavior and Personality: an international journal*, 32:361–372, January 2004. doi: 10.2224/sbp.2004.32.4.361.
- [160] Richard A. Dienstbier. Attraction increases and decreases as a function of emotion-attribution and appropriate social cues. *Motivation and Emotion*, 3(2):201–218, June 1979. ISSN 0146-7239, 1573-6644. doi: 10.1007/BF01650604. URL <http://link.springer.com/10.1007/BF01650604>.
- [161] Adam Kendon. Movement coordination in social interaction: Some examples described. *Acta Psychologica*, 32(C):101–125, 1970. ISSN 00016918. doi: 10.1016/0001-6918(70)90094-6.
- [162] Ishabel M. Vicaria and Leah Dickens. Meta-Analyses of the Intra- and Interpersonal Outcomes of Interpersonal Coordination. *Journal of Nonverbal Behavior*, 40(4):335–361, 2016. ISSN 01915886. doi: 10.1007/s10919-016-0238-8. Publisher: Springer US.
- [163] Tanya L. Chartrand and Jessica L. Lakin. The Antecedents and Consequences of Human Behavioral Mimicry. *Annual Review of Psychology*, 64(1):285–308, 2013. ISSN 0066-4308. doi: 10.1146/annurev-psych-113011-143754.
- [164] Frank J. Bernieri and Robert Rosenthal. Interpersonal coordination: Behavior matching and interactional synchrony. In *Fundamentals of nonverbal behavior*, pages 401–432. Cambridge University Press, 1977.
- [165] Alexandra Paxton and Rick Dale. Interpersonal movement synchrony responds to high- and low-level conversational constraints. *Frontiers in Psychology*, 8(JUL):1–16, 2017. ISSN 16641078. doi: 10.3389/fpsyg.2017.01135.
- [166] Eva G. Krumhuber, Katja U. Likowski, and Peter Weyers. Facial Mimicry of Spontaneous and Deliberate Duchenne and Non-Duchenne Smiles. *Journal of Nonverbal Behavior*, 38(1):1–11, 2014. ISSN 01915886. doi: 10.1007/s10919-013-0167-8. ISBN: 1091901301678.
- [167] Harry Griffin, Giovanna Varni, Gualtiero Volpe, Gisela Tome, Maurizio Mancini, and Nadia Bianchi-Berthouze. Gesture mimicry in expression of laughter. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.
- [168] Tanya L. Chartrand, William W. Maddux, and Jessica L. Lakin. Beyond the Perception-Behavior Link: The Ubiquitous Utility and Motivational Moderators of Nonconscious Mimicry. In Ran R. Hassin, James S. Uleman, and John A. Bargh, editors, *The new unconscious*, pages 334–361. Oxford University Press, 2012. ISBN 978-0-19-984748-8. doi: 10.1093/acprof:oso/9780195307696.003.0014.

- [169] Agneta Fischer and Ursula Hess. Mimicking emotions. *Current Opinion in Psychology*, 17:151–155, 2017. ISSN 2352250X. doi: 10.1016/j.copsyc.2017.07.008. URL <http://dx.doi.org/10.1016/j.copsyc.2017.07.008>. Publisher: Elsevier Ltd.
- [170] Bo Xiao, Panayiotis Georgiou, Brian Baucom, and Shrikanth S. Narayanan. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, 17(7):1107–1119, 2015. ISSN 15209210. doi: 10.1109/TMM.2015.2432671.
- [171] Cynthia L. Crown. Coordinated Interpersonal Timing of Vision and Voice as a Function of interpersonal Attraction. *Journal of Language and Social Psychology*, 10(1):29–46, March 1991. ISSN 0261-927X. doi: 10.1177/0261927X91101002. URL <https://doi.org/10.1177/0261927X91101002>. Publisher: SAGE Publications Inc.
- [172] Young Yun Kim. Achieving synchrony: A foundational dimension of intercultural communication competence. *International Journal of Intercultural Relations*, 48:27–37, 2015. ISSN 01471767. doi: 10.1016/j.ijintrel.2015.03.016. URL <http://dx.doi.org/10.1016/j.ijintrel.2015.03.016>. Publisher: Elsevier Ltd.
- [173] Daniël Lakens. Movement synchrony and perceived entitativity. *Journal of Experimental Social Psychology*, 46(5):701–708, 2010. ISSN 00221031. doi: 10.1016/j.jesp.2010.03.015.
- [174] Sander L. Koole and Wolfgang Tschacher. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in Psychology*, 7 (June):1–17, 2016. ISSN 16641078. doi: 10.3389/fpsyg.2016.00862.
- [175] Fabian Ramseyer and Wolfgang Tschacher. SYNCHRONY IN DYADIC PSYCHOTHERAPY SESSIONS. *Simultaneity: Temporal Structures and Observer perspectives*, pages 329–347, 2008. URL <papers3://publication/uuid/EFF6F812-94D3-4951-8DCD-6CD74E408EA0>.
- [176] Hayley Hung and Daniel Gatica-Perez. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, October 2010. ISSN 1941-0077. doi: 10.1109/TMM.2010.2055233. Conference Name: IEEE Transactions on Multimedia.
- [177] Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44, 2011. ISSN 09567976. doi: 10.1177/0956797610392928. ISBN: 1467-9280 (Electronic)\r0956-7976 (Linking).
- [178] E Prochazkova, E E Sjak-Shie, F Behrens, D Lindh, M E Kret, and * Affiliations. The choreography of human attraction: physiological synchrony in a blind date setting, 2019. URL <http://dx.doi.org/10.1101/748707>. Pages: 1-33.
- [179] Wolfgang Tschacher and Deborah Meier. Physiological synchrony in psychotherapy sessions. *Psychotherapy Research*, 0(0):1–16, 2019. ISSN 14684381. doi: 10.1080/10503307.2019.1612114. URL <https://doi.org/10.1080/10503307.2019.1612114>. Publisher: Taylor & Francis.

- [180] Andrew Chang, Haley E Kragness, Wei Tsou, Dan J Bosnyak, Anja Thiede, and Laurel J Trainor. Body sway predicts romantic interest in speed dating. *Social Cognitive and Affective Neuroscience*, 16(1-2):185–192, January 2021. ISSN 1749-5016, 1749-5024. doi: 10.1093/scan/nsaa093. URL <https://academic.oup.com/scan/article/16/1-2/185/5873242>.
- [181] Sebastian Feese, Bert Arnrich, Gerhard Troster, Bertolt Meyer, and Klaus Jonas. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pages 520–525, 2012. ISBN 978-0-7695-4848-7. doi: 10.1109/SocialCom-PASSAT.2012.48. Issue: September.
- [182] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern Recognition Letters*, 66:52–61, 2015. ISSN 01678655. doi: 10.1016/j.patrec.2015.03.005. URL <http://dx.doi.org/10.1016/j.patrec.2015.03.005>. Publisher: Elsevier Ltd. ISBN: 1090-0233.
- [183] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic. A multimodal database for mimicry analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6974 LNCS(PART 1):367–376, 2011. ISSN 03029743. doi: 10.1007/978-3-642-24600-5_40. ISBN: 9783642245992.
- [184] Rivka Levitan. *Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue*. PhD thesis, Columbia University, 2014.
- [185] Jens Edlund, Mattias Heldner, and Julia Hirschberg. Pause and gap length in face-to-face interaction. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2779–2782, 2009. Issue: January ISSN: 19909772.
- [186] Jan Michalsky and Heike Schoormann. Pitch convergence as an effect of perceived attractiveness and likability. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017-Augus, pages 2253–2256, 2017. doi: 10.21437/Interspeech.2017-1520. Issue: August ISSN: 19909772.
- [187] Jan Michalsky, Heike Schoormann, and Oliver Niebuhr. Conversational quality is affected by and reflected in prosodic entrainment. *Proceedings of the International Conference on Speech Prosody*, 2018-June(June):389–392, 2018. ISSN 23332042. doi: 10.21437/SpeechProsody.2018-79.
- [188] J Trouvain and K P Truong. Convergence of laughter in conversational speech: effects of quantity, temporal alignment and imitation. *Proceedings of the International Symposium on Imitation and Convergence in Speech, ISICS 2012, Aix-en-Provence, France*, 231287(231287), 2012.

- [189] Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts. Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, 66(1): 249–264, 2012. ISSN 0749596X. doi: 10.1016/j.jml.2011.07.004. URL <http://dx.doi.org/10.1016/j.jml.2011.07.004>. Publisher: Elsevier Inc. ISBN: 0749596X.
- [190] Taiki Ogata, Naoki Higo, Takayuki Nozawa, Eisuke Ono, Kazuo Yano, Koji Ara, and Yoshihiro Miyake. Interpersonal coevolution of body movements in daily face-to-face communication. *IEICE Transactions on Information and Systems*, E100D(10): 2547–2555, 2017. ISSN 17451361. doi: 10.1587/transinf.2016EDP7444.
- [191] Naoki Higo, Ken Ichiro Ogawa, Juichi Minemura, Bujie Xu, Takayuki Nozawa, Taiki Ogata, Koji Ara, Kazuo Yano, and Yoshihiro Miyake. Interpersonal similarity between body movements in face-to-face communication in daily life. *PLoS ONE*, 9(7):1–10, 2014. ISSN 19326203. doi: 10.1371/journal.pone.0102019.
- [192] Sarah Weidman, Mara Breen, and Katherine C. Haydon. Prosodic speech entrainment in romantic relationships. *Proceedings of the International Conference on Speech Prosody*, 2016-Janua(August):508–512, 2016. ISSN 23332042. doi: 10.21437/speechprosody.2016-104.
- [193] Rivk Levitan, Stefan Benus, Agustín Gravano, and Juli Hirschberg. Entrainment and turn-taking in human-human dialogue. *AAAI Spring Symposium - Technical Report*, SS-15-07(Gravano 2009):44–51, 2015. ISBN: 9781577357117.
- [194] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 11–19, 2012. ISBN: 1937284204.
- [195] Jessica Phillips-silver, c. athena Aktipis, and gregory a. Bryant. The ecology of entrainment: foundations of coordinated rhythmic movement. *Music Perception*, 28(09):3–14, 2010.
- [196] Linda Tickle-Degnen and Robert Rosenthal. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry - PSYCHOL INQ*, 1:285–293, October 1990. doi: 10.1207/s15327965pli0104_1.
- [197] Shanhong Luo and Guangjian Zhang. What leads to romantic attraction: Similarity, reciprocity, security, or beauty? Evidence from a speed-dating study. *Journal of Personality*, 77(4):933–963, 2009. ISSN 00223506. doi: 10.1111/j.1467-6494.2009.00570.x. ISBN: 1467-6494.
- [198] Ekin Gedik and Hayley Hung. Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):163:1–163:24, 2018. doi: 10.1145/3287041. URL <http://doi.org/10.1145/3287041>.

- [199] Yuval Hart, Efrat Czerniak, Orit Karnieli-Miller, Avraham E. Mayo, Amitai Ziv, Anat Biegon, Atay Citron, and Uri Alon. Automated Video Analysis of Non-verbal Communication in a Medical Setting. *Frontiers in Psychology*, 7, 2016. ISSN 1664-1078. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01130>.
- [200] Andrea Stevenson Won, Jeremy N. Bailenson, and Joris H. Janssen. Automatic Detection of Nonverbal Behavior Predicts Learning in Dyadic Interactions. *IEEE Transactions on Affective Computing*, 5(2):112–125, April 2014. ISSN 1949-3045. doi: 10.1109/TAFFC.2014.2329304. Conference Name: IEEE Transactions on Affective Computing.
- [201] Andrea Won, Jeremy Bailenson, Suzanne Stathatos, and Dai Wenqing. Automatically Detected Nonverbal Behavior Predicts Creativity in Collaborating Dyads. *Journal of Nonverbal Behavior*, 38, September 2014. doi: 10.1007/s10919-014-0186-0.
- [202] Claudio Martella, Ekin Gedik, Laura Cabrera-quiros, Gwenn Englebienne, Hayley Hung, Instituto Tecnológico, De Costa Rica, and Costa Rica. How Was It ? Exploiting Smartphone Sensing to Measure Implicit Audience Responses to Live Performances Categories and Subject Descriptors. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 201–210, 2015. ISBN 978-1-4503-3459-4. doi: 0.1145/2733373.2806276.
- [203] Claude Nadeau and Yoshua Bengio. Inference for the Generalization Error. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL <https://papers.nips.cc/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-Abstract.html>.
- [204] Remco R. Bouckaert and Eibe Frank. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056, pages 3–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-22064-0 978-3-540-24775-3. doi: 10.1007/978-3-540-24775-3_3. URL http://link.springer.com/10.1007/978-3-540-24775-3_3. Series Title: Lecture Notes in Computer Science.
- [205] Norene Kelly and Stephen Gilbert. The WEAR Scale: Developing a Measure of the Social Acceptability of a Wearable Device. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 2864–2871, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2892331. URL <http://doi.org/10.1145/2851581.2892331>.
- [206] Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 2009. ISSN 04194209. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.3741>.

- [207] David McNeill. Hand and Mind: What Gestures Reveal About Thought. *University of Chicago Press*, 1994. ISSN 00238309. doi: 10.1177/002383099403700208.
- [208] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *British Machine Vision Conference 2018, BMVC 2018*, pages 1–12, 2019. arXiv: 1802.00977.
- [209] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao Shu Fang, and Cewu Lu. Crowd-pose: Efficient crowded scenes pose estimation and a new benchmark. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10855–10864, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01112. arXiv: 1812.00324v2 ISBN: 9781728132938.
- [210] Hayley Hung, Gwenn Englebienne, and Laura Cabrera Quiros. Detecting conversing groups with a single worn accelerometer. *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, pages 84–91, 2014. doi: 10.1145/2663204.2663228. URL <http://dl.acm.org/citation.cfm?doid=2663204.2663228>. ISBN: 9781450328852.
- [211] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng Lin Liu. Action recognition by dense trajectories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011. ISBN 978-1-4577-0394-2. doi: 10.1109/CVPR.2011.5995407. arXiv: 1505.04868 ISSN: 10636919.
- [212] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2014. ISSN 1090235X. doi: 10.1016/j.cviu.2016.03.013. arXiv: 1405.4506.
- [213] Dan Oneata, Jakob Verbeek, and Cordelia Schmid Inria. Efficient Action Localization with Approximately Normalized Fisher Vectors. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [214] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *European Conference on Computer Vision (ECCV)*, 2010. ISSN: 10603271.
- [215] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image Classification with the Fisher Vector : Theory and Practice. *International journal of computer vision*, 105(3):222–245, 2013. ISSN 0021-924X (Print). doi: 10.1007/s11263-013-0636-x. ISBN: 0920-5691.
- [216] Heng Wang, Cordelia Schmid, Heng Wang, Cordelia Schmid, Action Recognition, Trajectories Iccv, Heng Wang, and Cordelia Schmid. Action Recognition with Improved Trajectories. *ICCV - IEEE International Conference on Computer Vision*, December:3551–3558, 2013.
- [217] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *NIPS Proceedings*, 2014. ISSN 1098-6596. doi: 10.

- 1017/CBO9781107415324.004. URL <http://www.ncbi.nlm.nih.gov/pubmed/16952995>. arXiv: 1406.2199v1 ISBN: 9788578110796.
- [218] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.59.
- [219] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:4489–4497, 2015. ISSN 15505499. doi: 10.1109/ICCV.2015.510. arXiv: 1412.0767 ISBN: 9781467383912.
- [220] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.502. arXiv: 1705.07750 ISSN: 0032082X.
- [221] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Lecun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00675. arXiv: 1711.11248 ISBN: 9781538664209.
- [222] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *arXiv:1604.01753 [cs]*, July 2016. URL <http://arxiv.org/abs/1604.01753>. arXiv: 1604.01753.
- [223] Gunnar A. Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What Actions are Needed for Understanding Human Actions in Videos? *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2156–2165, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.235. ISBN: 9781538610329.
- [224] Guilhem Cheron, Ivan Laptev, and Cordelia Schmid. P-CNN: Pose-based CNN features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.368. arXiv: 1506.03607 ISSN: 15505499.
- [225] Leonid Pishchulin, Mykhaylo Andriluka, and Bernt Schiele. Fine-grained activity recognition with holistic and pose based features. In *Pattern Recognition: 36th German Conference*, volume 8753, pages 678–689, 2014. ISBN 978-3-319-11751-5. doi: 10.1007/978-3-319-11752-2_56. arXiv: 1406.1881 ISSN: 16113349.
- [226] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. Voice activity detection by upper body motion analysis and unsupervised domain adaptation. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 1260–1269, 2019. doi: 10.1109/ICCVW.2019.00159. ISBN: 9781728150239.

- [227] Cigdem Beyan, Muhammad Shahid, and Vittorio Murino. RealVAD: A Real-world Dataset and A Method for Voice Activity Detection by Body Motion Analysis. *x*, 9210(c):1–16, 2020. doi: 10.1109/tmm.2020.3007350.
- [228] Yongtao Hu, Jimmy Sj Ren, Jingwen Dai, Chang Yuan, Li Xu, and Wenping Wang. Deep Multimodal Speaker Naming. In *MM '15: Proceedings of the 23rd ACM international conference on Multimedia*, pages 1107–1110, Brisbane, Australia, 2015. Association for Computing Machinery. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806293. URL <https://doi.org/10.1145/2733373.2806293>. arXiv: 1507.04831 ISBN: 9781450334594.
- [229] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496, May 2020. doi: 10.1109/ICASSP40776.2020.9053900. URL <https://ieeexplore.ieee.org/document/9053900>. ISSN: 2379-190X.
- [230] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A Survey. *Pattern Recognition Letters*, 0:1–9, 2018. ISSN 01678655. doi: 10.1016/j.patrec.2018.02.010. URL <https://doi.org/10.1016/j.patrec.2018.02.010>. arXiv: 1707.03502 Publisher: Elsevier B.V. ISBN: 01678655 (ISSN).
- [231] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.471. ISBN: 9781479951178.
- [232] Anna Esposito and Antonietta M. Esposito. On speech and gestures synchrony. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6800 LNCS:252–272, 2011. ISSN 03029743. doi: 10.1007/978-3-642-25775-9_25. ISBN: 9783642257742.
- [233] Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5650–5659, 2017. doi: 10.1109/CVPR.2017.599. arXiv: 1612.06371 ISBN: 9781538604571.
- [234] Yuqing Chen and Yang Xue. A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 5–9. IEEE, 2015. ISBN 978-1-4799-8697-2. doi: 10.1109/SMC.2015.263.
- [235] Hristijan Gjoreski, Jani Bizjak, Martin Gjoreski, and Matjaž Gams. Comparing Deep and Classical Machine Learning Methods for Human Activity Recognition using Wrist Accelerometer, 2016. URL [https://sites.cc.gatech.edu/~alanwags/DLAI2016/2.%20\(Gjoreski+\)%20Comparing%20Deep%20and%20Classical%20Machine%](https://sites.cc.gatech.edu/~alanwags/DLAI2016/2.%20(Gjoreski+)%20Comparing%20Deep%20and%20Classical%20Machine%20)

20Learning%20Methods%20for%20Human%20Activity%20Recognition%20using%20Wrist%20Accelerometer.pdf.

- [236] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS Proceedings*, 2012. doi: 10.1201/9781420010749. ISBN: 9781420010749.
- [237] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. The second dihard diarization challenge: Dataset, task, and baselines. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-Sept:978–982, 2019. ISSN 19909772. doi: 10.21437/Interspeech.2019-1268. arXiv: 1906.07839v1.
- [238] Herve Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2020*, 2020. doi: 10.1109/icassp40776.2020.9052974. arXiv: 1911.01255v1.
- [239] Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Leibny Paola Garcia-Perera. End-to-end Domain-Adversarial Voice Activity Detection. In *Proc. Interspeech 2020*, pages 3685–3689, 2020. doi: 10.21437/Interspeech.2020-2285. URL <http://arxiv.org/abs/1910.10655>. arXiv: 1910.10655.
- [240] Zheng Hua Tan, Achintya kr Sarkar, and Najim Dehak. rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech and Language*, 59:1–21, 2020. ISSN 10958363. doi: 10.1016/j.csl.2019.06.005.
- [241] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalic, Daniel Balliet, and Hayley Hung. Recognizing Perceived Interdependence in Face-to-Face Negotiations through Multimodal Analysis of Nonverbal Behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21*, pages 121–130, New York, NY, USA, October 2021. Association for Computing Machinery. ISBN 978-1-4503-8481-0. doi: 10.1145/3462244.3479935. URL <http://doi.org/10.1145/3462244.3479935>.
- [242] William Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, 75(4):825–861, August 2007. ISSN 0022-3506. doi: 10.1111/j.1467-6494.2007.00458.x.
- [243] Jennifer G. La Guardia and Richard M. Ryan. Why identities fluctuate: variability in traits as a function of situational variations in autonomy support. *Journal of Personality*, 75(6):1205–1228, December 2007. ISSN 0022-3506. doi: 10.1111/j.1467-6494.2007.00473.x.
- [244] Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. Nonverbal Communication. *Annual Review of Psychology*, 70:271–294, January 2019. ISSN 1545-2085. doi: 10.1146/annurev-psych-010418-103145.

- [245] Katherine Osborne-Crowley. Social Cognition in the Real World: Reconnecting the Study of Social Cognition With Social Reality. *Review of General Psychology*, 24(2):144–158, June 2020. ISSN 1089-2680. doi: 10.1177/1089268020906483. URL <https://doi.org/10.1177/1089268020906483>. Publisher: SAGE Publications Inc.
- [246] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning Delicate Local Representations for Multi-person Pose Estimation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pages 455–472, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58579-2. doi: 10.1007/978-3-030-58580-8_27. URL http://doi.org/10.1007/978-3-030-58580-8_27.
- [247] Chittaranjan Andrade. Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation. *Indian Journal of Psychological Medicine*, 40(5):498–499, 2018. ISSN 0253-7176. doi: 10.4103/IJPSYM.IJPSYM_334_18. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6149308/>.
- [248] Elise L. LeMoyne, François Courtemanche, Marc Fredette, and Pierre-Majorique Léger. How wild is too wild: Lessons learned and recommendations for ecological validity in physiological computing research. In *International Conference on Physiological Computing Systems*, pages 123–130, January 2018. doi: 10.5220/0006962901230130. URL <https://api.semanticscholar.org/CorpusID:52899722>.
- [249] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. Chapter 11 - Complex conversational scene analysis using wearable sensors. In Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe, editors, *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 225–245. Academic Press, January 2019. ISBN 978-0-12-814601-9. doi: 10.1016/B978-0-12-814601-9.00019-5. URL <https://www.sciencedirect.com/science/article/pii/B9780128146019000195>.
- [250] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, MPVA '10, pages 37–42, New York, NY, USA, October 2010. Association for Computing Machinery. ISBN 978-1-4503-0167-1. doi: 10.1145/1878039.1878048. URL <http://doi.org/10.1145/1878039.1878048>.
- [251] Madhumita Murgia. Who’s using your face? The ugly truth about facial recognition. *Financial Times*, 2019.
- [252] Nicolò Carissimi, Paolo Rota, Cigdem Beyan, and Vittorio Murino. Filling the Gaps: Predicting Missing Joints of Human Poses Using Denoising Autoencoders. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11130, pages 364–379. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11011-6 978-3-030-11012-3. doi: 10.1007/978-3-030-11012-3_29. URL http://link.springer.com/10.1007/978-3-030-11012-3_29. Series Title: Lecture Notes in Computer Science.

- [253] C.A. Raman, S. Tan, and H.S. Hung. A Modular Approach for Synchronized Wireless Multimodal Multisensor Data Acquisition in Highly Dynamic Social Settings: 28th ACM International Conference on Multimedia. *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3586–3594, 2020. doi: 10.1145/3394171.3413697. URL <http://www.scopus.com/inward/record.url?scp=85106924663&partnerID=8YFLogxK>.
- [254] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision?, July 2020. URL <http://arxiv.org/abs/2006.16923>. arXiv:2006.16923 [cs, stat].
- [255] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. Improving Social Awareness Through DANTE: Deep Affinity Network for Clustering Conversational Interactants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):20:1–20:23, May 2020. doi: 10.1145/3392824. URL <http://doi.org/10.1145/3392824>.
- [256] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, volume 9007, pages 658–675. Springer International Publishing, Cham, 2015. ISBN 978-3-319-16813-5 978-3-319-16814-2. doi: 10.1007/978-3-319-16814-2_43. URL http://link.springer.com/10.1007/978-3-319-16814-2_43. Series Title: Lecture Notes in Computer Science.
- [257] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *arXiv:1612.03153 [cs]*, December 2016. URL <http://arxiv.org/abs/1612.03153>. arXiv: 1612.03153.
- [258] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling up Crowd-sourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013. ISSN 0920-5691. doi: 10.1007/s11263-012-0564-1.
- [259] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Buló, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter: 4660–4668, 2015. ISSN 15505499. doi: 10.1109/ICCV.2015.529. ISBN: 9781467383912.
- [260] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013. ISSN 1468-0394. doi: 10.1111/j.1468-0394.2012.00622.x. URL <http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0394.2012.00622.x>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0394.2012.00622.x>.

- [261] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLOS ONE*, 5(7):e11596, July 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0011596. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011596>. Publisher: Public Library of Science.
- [262] Marion Hoffman, Per Block, Timon Elmer, and Christoph Stadtfeld. A model for the dynamics of face-to-face interactions in social groups. *Network Science*, 8(S1):S4–S25, July 2020. ISSN 2050-1242, 2050-1250. doi: 10.1017/nws.2020.3. URL <http://www.cambridge.org/core/journals/network-science/article/model-for-the-dynamics-of-facetoface-interactions-in-social-groups/5EE32074370B3C443EE9AA2519602338>. Publisher: Cambridge University Press.
- [263] Martin Atzmueller and Florian Lemmerich. Homophily at Academic Conferences. In *Companion Proceedings of the The Web Conference 2018, WWW '18*, pages 109–110, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3186953. URL <http://doi.org/10.1145/3184558.3186953>.
- [264] Daniel Olguín Olguín, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):43–55, 2009. ISSN 10834419. doi: 10.1109/TSMCB.2008.2006638.
- [265] Daniel Chaffin, Ralph Heidl, John R. Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. The Promise and Perils of Wearable Sensors in Organizational Research. *Organizational Research Methods*, 20(1):3–31, January 2017. ISSN 1094-4281. doi: 10.1177/1094428115617004. URL <https://doi.org/10.1177/1094428115617004>. Publisher: SAGE Publications Inc.
- [266] Alessio Rosatelli, Ekin Gedik, and Hayley Hung. Detecting F-formations & Roles in Crowded Social Scenes with Wearables: Combining Proxemics & Dynamics using LSTMs. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 147–153, September 2019. doi: 10.1109/ACIIW.2019.8925179.
- [267] Stephanie Tan, David M. J. Tax, and Hayley Hung. Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–22, March 2021. ISSN 2474-9567. doi: 10.1145/3448122. URL <https://dl.acm.org/doi/10.1145/3448122>.
- [268] University of York. University of York Research Data Management, 2021. URL <https://www.york.ac.uk/library/info-for/researchers/data/sharing/access/>.
- [269] Utrecht University. Utrecht university research data management, 2021. URL <https://www.uu.nl/en/research/research-data-management/guides/handling-personal-data>.

- [270] GoPro. Go pro hero 7 black, 2018. URL <https://gopro.com/en/nl/shop/cameras/hero7-black/CHDX-701-master.html>.
- [271] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, pages 1–8, 2013. doi: 10.1109/FG.2013.6553805. Publisher: IEEE ISBN: 9781467355452.
- [272] Laura Cabrera-Quiros and Hayley Hung. Who is where? Matching People in Video to Wearable Acceleration During Crowded Mingling Events. In *Proceedings of the 24th ACM international conference on Multimedia, MM '16*, pages 267–271, New York, NY, USA, October 2016. Association for Computing Machinery. ISBN 978-1-4503-3603-1. doi: 10.1145/2964284.2967224. URL <http://doi.org/10.1145/2964284.2967224>.
- [273] Laura Cabrera-Quiros and Hayley Hung. A Hierarchical Approach for Associating Body-Worn Sensors to Video Regions in Crowded Mingling Scenarios. *IEEE Transactions on Multimedia*, 21(7):1867–1879, July 2019. ISSN 1941-0077. doi: 10.1109/TMM.2018.2888798. Conference Name: IEEE Transactions on Multimedia.
- [274] Hayley Hung, Chirag Raman, Ekin Gedik, Stephanie Tan, and Jose Vargas Quiros. Multimodal Data Collection for Social Interaction Analysis In-the-Wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2714–2715, Nice France, October 2019. ACM. ISBN 978-1-4503-6889-6. doi: 10.1145/3343031.3351320. URL <https://dl.acm.org/doi/10.1145/3343031.3351320>.
- [275] CVAT Company. Computer Vision Annotation Tool (CVAT), 2019. URL <https://github.com/openvinotoolkit/cvat>.
- [276] Daniel Gatica-Perez. Analyzing Group Interactions in Conversations: a Review. In *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 41–46, September 2006. doi: 10.1109/MFI.2006.265658.
- [277] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164, Tokyo Japan, March 2018. ACM. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172969. URL <https://dl.acm.org/doi/10.1145/3172944.3172969>.
- [278] Chirag Raman and Hayley Hung. Towards automatic estimation of conversation floors within F-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 175–181, Cambridge, United Kingdom, September 2019. doi: 10.1109/ACIIW.2019.8925065. URL <https://doi.ieeecomputersociety.org/10.1109/ACIIW.2019.8925065>.
- [279] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860, May 2011. ISSN 1558-7924.

- doi: 10.1109/TASL.2010.2066267. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.
- [280] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.
- [281] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, October 2017. doi: 10.1109/ICCV.2017.322. ISSN: 2380-7504.
- [282] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019. URL <https://github.com/facebookresearch/detectron2>.
- [283] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February 2015. URL <http://arxiv.org/abs/1405.0312>. arXiv: 1405.0312.
- [284] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo Vadis, Skeleton Action Recognition ? *arXiv:2007.02072 [cs]*, July 2020. URL <http://arxiv.org/abs/2007.02072>. arXiv: 2007.02072 version: 1.
- [285] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:1578–1585, 2017. doi: 10.1109/IJCNN.2017.7966039. arXiv: 1611.06455 ISBN: 9781509061815.
- [286] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 248–257, August 2021. doi: 10.1145/3447548.3467231. URL <http://arxiv.org/abs/2012.08791>. arXiv:2012.08791 [cs, stat].
- [287] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2798607. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [288] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. *arXiv:2003.14111 [cs]*, May 2020. URL <http://arxiv.org/abs/2003.14111>. arXiv: 2003.14111.
- [289] Marco Cristani, R. Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A Social Signal Processing perspective. *Neurocomputing*, 100:86–97, January 2013. ISSN 0925-2312. doi: 10.1016/j.neucom.2011.12.038. URL <https://www.sciencedirect.com/science/article/pii/S0925231212003141>.

- [290] Chirag Raman, Navin Raj Prabhu, and Hayley Hung. Perceived Conversation Quality in Spontaneous Interactions, July 2022. URL <http://arxiv.org/abs/2207.05791>. arXiv:2207.05791 [cs].
- [291] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces, ICMI '07*, pages 271–278, New York, NY, USA, November 2007. Association for Computing Machinery. ISBN 978-1-59593-817-6. doi: 10.1145/1322192.1322239. URL <http://doi.org/10.1145/1322192.1322239>.
- [292] Julia Eberle, Karsten Stegmann, Alain Barrat, Frank Fischer, and Kristine Lund. Initiating scientific collaborations across career levels and disciplines – a network analysis on behavioral data. *International Journal of Computer-Supported Collaborative Learning*, 16(2):151–184, June 2021. ISSN 1556-1615. doi: 10.1007/s11412-021-09345-7. URL <https://doi.org/10.1007/s11412-021-09345-7>.
- [293] Nigel Pleasants. Free Will, Determinism and the “Problem” of Structure and Agency in the Social Sciences. *Philosophy of the Social Sciences*, 49(1):3–30, January 2019. ISSN 0048-3931. doi: 10.1177/0048393118814952. URL <https://doi.org/10.1177/0048393118814952>. Publisher: SAGE Publications Inc.
- [294] Chirag Raman, Hayley Hung, and Marco Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues, August 2022. URL <http://arxiv.org/abs/2107.13576>. arXiv:2107.13576 [cs].
- [295] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, December 2021. URL <http://arxiv.org/abs/1803.09010>. arXiv:1803.09010.
- [296] German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn’t see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 139–178. PMLR, March 2022. URL <https://proceedings.mlr.press/v173/barquero22b.html>. ISSN: 2640-3498.
- [297] Chirag Raman, Hayley Hung, and Marco Loog. Why Did This Model Forecast This Future? Closed-Form Temporal Saliency Towards Causal Explanations of Probabilistic Forecasts, June 2022. URL <http://arxiv.org/abs/2206.00679>. arXiv:2206.00679 [cs, math].
- [298] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. Defining and Quantifying Conversation Quality in Spontaneous Interactions. *arXiv:2009.12842 [cs]*, September 2020. URL <http://arxiv.org/abs/2009.12842>. arXiv: 2009.12842.
- [299] OpenCV. Open source computer vision library, 2015. URL <https://github.com/opencv/opencv>.

- [300] IDIAP. IDIAP multi camera calibration suite, 2017. URL <https://github.com/idiap/multicamera-calibration>.
- [301] TDK. TDKICM20948, 2024. URL <https://invensense.tdk.com/products/motion-tracking/9-axis/icm-20948/>.
- [302] Sileye O. Ba and Jean-Marc Odobez. Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, January 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2010.69. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [303] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on Multi-Stage Networks for Human Pose Estimation, May 2019. URL <http://arxiv.org/abs/1901.00148>. arXiv:1901.00148 [cs].
- [304] Bowen Cheng, Bin Xiao, Jingdong Wang, Humphrey Shi, and Lei Zhang. HigherHR-Net: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, June 2020. doi: 10.1109/CVPR42600.2020.00543.
- [305] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://papers.nips.cc/paper/2017/hash/8edd72158ccd2a879f79cb2538568fdc-Abstract.html>.
- [306] Ignacio Oguiza. tsai - A state-of-the-art deep learning library for time series and sequential data, 2022. URL <https://github.com/timeseriesAI/tsai>.
- [307] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 216–222, October 2003. doi: 10.1109/ISWC.2003.1241414. ISSN: 1530-0811.
- [308] Oren Lederman, Dan Calacci, Angus MacMullen, Daniel C. Fehder, Fiona E. Murray, and Alex ‘Sandy’ Pentland. Open Badges: A Low-Cost Toolkit for Measuring Team Communication and Dynamics, 2017. URL <http://arxiv.org/abs/1710.01842>. arXiv: 1710.01842.
- [309] Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild. *IEEE Transactions on Affective Computing*, pages 1–17, 2023. ISSN 1949-3045. doi: 10.1109/TAFFC.2023.3269003. URL <https://ieeexplore.ieee.org/document/10136533>. Conference Name: IEEE Transactions on Affective Computing.
- [310] Yu Kong and Yun Fu. Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, 130(5):1366–1401, May 2022. ISSN 1573-1405. doi: 10.1007/s11263-022-01594-9. URL <https://doi.org/10.1007/s11263-022-01594-9>.

- [311] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Multi-task Learning of Social Psychology Assessments and Nonverbal Features for Automatic Leadership Identification. In *ICMI*, pages 451–455, 2017. ISBN 978-1-4503-5543-8. doi: 10.1145/3136755.3136812. URL <https://doi.org/10.1145/3136755.3136812>. Issue: 5.
- [312] Laura Cabrera-quiros, Ekin Gedik, and Hayley Hung. Transductive Parameter Transfer, Bags of Dense Trajectories and MILES for No-Audio Multimodal Speech Detection. In *Working Notes Proceedings of the MediaEval 2018 Workshop*, 2018.
- [313] Marco Cristani, Anna Pesarin, Alessandro Vinciarelli, Marco Crocco, and Vittorio Murino. Look at who’s talking: Voice activity detection by automated gesture analysis. In Reiner Wichert, Kristof Van Laerhoven, and Jean Gelissen, editors, *Constructing ambient intelligence*, pages 72–80, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-31479-7.
- [314] Yixin Chen, Jinbo Bi, and James Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006. ISSN 01628828. doi: 10.1109/TPAMI.2006.248. Publisher: IEEE ISBN: 0162-8828 VO - 28.
- [315] Aleksandar Matic, Venet Osmani, and Oscar Mayora. Automatic Sensing of Speech Activity and Correlation with Mood Changes. In Subhas Chandra Mukhopadhyay and Octavian A. Postolache, editors, *Pervasive and Mobile Sensing and Computing for Healthcare: Technological and Social Issues*, Smart Sensors, Measurement and Instrumentation, pages 195–205. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-32538-0. doi: 10.1007/978-3-642-32538-0_9. URL https://doi.org/10.1007/978-3-642-32538-0_9.
- [316] Ekin Gedik, Laura Cabrera-Quiros, and Hayley Hung. No-Audio Multimodal Speech Detection task at MediaEval 2019. In *Working Notes Proceedings of the MediaEval 2019 Workshop*, page 3, 2019.
- [317] Jose Vargas-Quiros, Laura Cabrera-Quiros, and Hayley Hung. No-audio speaking status detection in crowded settings via visual pose-based filtering and wearable acceleration, November 2022. URL <http://arxiv.org/abs/2211.00549>. arXiv:2211.00549 [cs, eess].
- [318] Punarjay Chakravarty and Tinne Tuytelaars. Cross-modal supervision for learning active speaker detection in video. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9909 LNCS:285–301, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46454-1_18. arXiv: 1603.08907 ISBN: 9783319464534.
- [319] FFmpeg. ffmpeg tool, 2016. URL <http://ffmpeg.org/>.
- [320] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021*, 2021.

- [321] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending speech separation to noisy environments. In *Proc. Interspeech*, September 2019.
- [322] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Krizan, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, and others. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.
- [323] Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, Jitendra Malik, Ross Girshick, Matt Feiszli, Aaron Adcock, Wan-Yen Lo, and Christoph Feichtenhofer. PyTorchVideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM international conference on multimedia*, 2021.
- [324] Senko K. Maynard. Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, 11(5):589–606, October 1987. ISSN 0378-2166. doi: 10.1016/0378-2166(87)90181-0. URL <https://www.sciencedirect.com/science/article/pii/0378216687901810>.
- [325] G. McKeown, W. Curran, J. Wagner, F. Lingenfelter, and E. André. The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 166–172, September 2015. doi: 10.1109/ACII.2015.7344567.
- [326] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160, June 2011. doi: 10.1109/CVPR.2011.5995586. ISSN: 1063-6919.
- [327] Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton. Gtrace: General Trace Program Compatible with EmotionML. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 709–710, September 2013. doi: 10.1109/ACII.2013.126. ISSN: 2156-8111.
- [328] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 3–9, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133953. URL <http://doi.org/10.1145/3133944.3133953>.

- [329] Karan Sharma, Claudio Castellini, Egon L. van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific Data*, 6(1):196, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0209-0. URL <http://www.nature.com/articles/s41597-019-0209-0>.
- [330] Jeffrey M Girard. CARMA: Software for continuous affect rating and media annotation. 2(1):e5, 2014. doi: 10.5334/jors.ar.
- [331] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 723–732, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00081. URL <https://ieeexplore.ieee.org/document/9010921/>.
- [332] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-Aware Siamese Network for Human Pose Estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12365, pages 396–412. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58564-8 978-3-030-58565-5. doi: 10.1007/978-3-030-58565-5_24. URL https://link.springer.com/10.1007/978-3-030-58565-5_24. Series Title: Lecture Notes in Computer Science.
- [333] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, Tokyo Japan, October 2016. ACM. ISBN 978-1-4503-4556-9. doi: 10.1145/2993148.2993165. URL <https://dl.acm.org/doi/10.1145/2993148.2993165>.
- [334] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, January 2017. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2515617. Conference Name: IEEE Transactions on Affective Computing.
- [335] Reza Lotfian and Carlos Busso. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October 2019. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2736999. Conference Name: IEEE Transactions on Affective Computing.
- [336] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2016. ISSN 19493045. doi: 10.1109/TAFFC.2015.2493525. Publisher: IEEE.
- [337] James A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. Publisher: American Psychological Association.

- [338] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005. ISSN 0954-5794. doi: 10.1017/S0954579405050340. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367156/>.
- [339] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, January 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.15. Conference Name: IEEE Transactions on Affective Computing.
- [340] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, July 2015. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2392932. Conference Name: IEEE Transactions on Affective Computing.
- [341] Angeliki Metallinou and Shrikanth Narayanan. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 2013. doi: 10.1109/FG.2013.6553804. ISBN: 9781467355452.
- [342] json-schema org. JSON Schema, 2020. URL <https://json-schema.org/>.
- [343] Jose Vargas Quiros. covfee: continuous video feedback tool, 2020. URL <https://josedvq.github.io/covfee/docs>.
- [344] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000. tex.citeulike-article-id: 2236121 tex.posted-at: 2008-01-15 19:21:54 tex.priority: 4.
- [345] Khiet P Truong, Jurgen Trouvain, and Michel-pierre Jansen. Towards an annotation scheme for complex laughter in speech corpora. In *Proc. Interspeech 2019*, pages 529–533, 2019. doi: 10.21437/Interspeech.2019-1557.
- [346] Charles Darwin. Expression of the emotions in Man and Animals. *Nature*, 36(926): 294–295, 1887. ISSN 00280836. doi: 10.1038/036294c0. ISBN: 0195158067.
- [347] Judee K Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli. *Social Signal Processing*. Cambridge University Press, 2017. ISBN 978-1-107-16126-9.
- [348] Maurizio Mancini, Giovanna Varni, Donald Glowinski, and Gualtiero Volpe. Computing and evaluating the body laughter index. *Lecture Notes in Computer Science*, 7559 LNCS:90–98, 2012. ISBN: 9783642340130.
- [349] R Niewiadomski, J Urbain, C Pelachaud, and T Dutoit. Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases. *Proceedings of the 4th International Workshop on Corpora for Research on Emotion*, pages 25–32, 2012.

- [350] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Laughter Recognition Using Non-invasive Wearable Devices. *PervasiveHealth'19: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019. doi: 10.1145/3329189.3329216.
- [351] Kevin El Haddad, Sandeep Nallan Chakravarthula, and James Kennedy. Smile and Laugh Dynamics in Naturalistic Dyadic Interactions: Intensity Levels, Sequences and Roles. In *2019 International Conference on Multimodal Interaction, ICMI '19*, pages 259–263, New York, NY, USA, October 2019. Association for Computing Machinery. ISBN 978-1-4503-6860-5. doi: 10.1145/3340555.3353764. URL <https://doi.org/10.1145/3340555.3353764>.
- [352] Stavros Petridis and Maja Pantic. Audiovisual discrimination between laughter and speech. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5117–5120, 2008. ISSN 1520-6149. doi: 10.1109/ICASSP.2008.4518810.
- [353] Stavros Petridis, Brais Martinez, and Maja Pantic. The MAHNOB Laughter database. *Image and Vision Computing*, 31(2):186–202, 2013. ISSN 02628856.
- [354] Khiet P. Truong and David A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007. ISSN 01676393. doi: 10.1016/j.specom.2007.01.001. ISBN: 0167-6393.
- [355] Khiet P. Truong and Jürgen Trouvain. On the acoustics of overlapping laughter in conversational speech. *INTERSPEECH 2012*, 1:850–853, 2012.
- [356] Keith Oatley and P.N. Johnson-Laird. Cognitive approaches to emotions. *Trends in Cognitive Sciences*, 18(3):134–140, March 2014. ISSN 13646613. doi: 10.1016/j.tics.2013.12.004.
- [357] Klaus R. Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351, November 2009. ISSN 0269-9931. doi: 10.1080/02699930902928969.
- [358] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. Robust Laughter Detection in Noisy Environments. In *Interspeech 2021*, pages 2481–2485. ISCA, August 2021.
- [359] Stavros Petridis, Maelle Leveque, and Maja Pantic. Audiovisual detection of laughter in human-machine interaction. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 129–134, 2013. doi: 10.1109/ACII.2013.28. \.
- [360] Radoslaw Niewiadomski and Catherine Pelachaud. Towards Multimodal Expression of Laughter. In *Intelligent Virtual Agents*, volume 7502, page 6221, 2012. doi: 10.1007/978-3-642-33197-8.
- [361] Michel-pierre Jansen, Khiet P Truong, Deniece S Nazareth, and Dirk K J Heylen. Introducing MULAI : A Multimodal Database of Laughter during Dyadic Interactions.

- In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4333–4342, 2020.
- [362] Robert Provine. *Laughter : a scientific investigation*. Penguin Press, 2001.
- [363] Kevin El Haddad, Hüseyin Cakmak, and Thierry Dutoit. On Laughter Intensity Level: Analysis and Estimation. In *Laughter Workshop*, September 2018.
- [364] Robert R. Provine. Laughter Punctuates Speech: Linguistic, Social and Gender Contexts of Laughter. *Ethology*, 95(4):291–298, 1993.
- [365] Elizabeth Holt. The last laugh : Shared laughter and topic termination. *Journal of Pragmatics*, 42(6):1513–1525, 2010. ISSN 0378-2166. doi: 10.1016/j.pragma.2010.01.011. Publisher: Elsevier B.V.
- [366] Nick O’donnell-Trujillo and Katherine Adams. Heheh in conversation: Some coordinating accomplishments of laughter. *Western Journal of Speech Communication*, 47(2):175–191, 1983. ISSN 01936700. doi: 10.1080/10570318309374114.
- [367] KP Truong and DA Van Leeuwen. Automatic Detection of Laughter. *Interspeech*, pages 485–488, 2005. ISBN: 1855212986.
- [368] Stavros Petridis and Maja Pantic. Audiovisual Laughter Detection Based on Temporal Features. *Belgian/Netherlands Artificial Intelligence Conference*, pages 351–352, 2008. ISSN 15687805. doi: 10.1145/1452392.1452402.
- [369] Stavros Petridis and Maja Pantic. Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help. *Why and When Visual*, 13(2):216–234, 2011.
- [370] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00630. URL <https://ieeexplore.ieee.org/document/9008780/>.
- [371] K. E. Haddad, S. Dupont, J. Urbain, and T. Dutoit. Speech-laugh: An HMM-based approach for amused speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4939–4943, April 2015. doi: 10.1109/ICASSP.2015.7178910. ISSN: 2379-190X.
- [372] Stavros Petridis. A Short Introduction to Laughter, 2015. URL <https://ibug.doc.ic.ac.uk/media/uploads/documents/shortintrotolaughter.pdf>.
- [373] Jürgen Trouvain. Segmenting phonetic units in laughter. *Proc. ICPhS ’03*, pages 2793–2796, 2003. ISSN 1876346485. doi: ISBN1-876346-48-5. ISBN: 1876346485.
- [374] Timothy R. Jordan and Lily Abedipour. The importance of laughing in your face: Influences of visual laughter on auditory laughter perception. *Perception*, 39(9): 1283–1285, 2010. ISSN 03010066. doi: 10.1068/p6752.

- [375] Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. ISSN 1574020X. doi: 10.1007/s10579-007-9040-x.
- [376] Maurizio Mancini, Jennifer Hofmann, Tracey Platt, Gualtiero Volpe, Giovanna Varni, Donald Glowinski, Willibald Ruch, and Antonio Camurri. Towards automated full body detection of laughter driven by human expert annotation. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 757–762, 2013. doi: 10.1109/ACII.2013.140.
- [377] Thomas H. Crystal and Arthur S. House. Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88(1):101–112, July 1990. ISSN 0001-4966. doi: 10.1121/1.399955.
- [378] Prolific. Prolific, 2014. URL <https://www.prolific.co>.
- [379] Kevin A. Hallgren. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Gff*, 82(2):218–226, 2012. ISSN 20000863. doi: 10.1080/11035896009449194.
- [380] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
- [381] Khiet P. Truong, Ronald Poppe, Iwan De Kok, and Dirk Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. *INTERSPEECH 2011*, pages 2973–2976, 2011. ISSN 19909772.
- [382] Sophie Huhn, Miriam Axt, Hanns-Christian Gunga, Martina Anna Maggioni, Stephen Munga, David Obor, Ali Sié, Valentin Boudo, Aditi Bunker, Rainer Sauerborn, Till Bärnighausen, and Sandra Barteit. The Impact of Wearable Technologies in Health Research: Scoping Review. *JMIR mHealth and uHealth*, 10(1):e34384, January 2022. ISSN 2291-5222. doi: 10.2196/34384.
- [383] Jessica R Golbus, Nicole A Pescatore, Brahmajee K Nallamothu, Nirav Shah, and Sachin Kheterpal. Wearable device signals and home blood pressure data across age, sex, race, ethnicity, and clinical phenotypes in the Michigan Predictive Activity & Clinical Trajectories in Health (MIPACT) study: a prospective, community-based observational study. *The Lancet Digital Health*, 3(11):e707–e715, November 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00138-2. URL <https://www.sciencedirect.com/science/article/pii/S2589750021001382>.
- [384] Stuart J. Fairclough, Sarah Taylor, Alex V. Rowlands, Lynne M. Boddy, and Robert J. Noonan. Average acceleration and intensity gradient of primary school children and associations with indicators of health and well-being. *Journal of Sports Sciences*, 37(18):2159–2167, September 2019. ISSN 0264-0414. doi: 10.1080/02640414.2019.1624313. URL <https://doi.org/10.1080/02640414.2019.1624313>. Publisher: Routledge _eprint: <https://doi.org/10.1080/02640414.2019.1624313>.

- [385] William V. Massey, Megan B. Stellino, and Margaret Fraser. Individual and environmental correlates of school-based recess engagement. *Preventive Medicine Reports*, 11:247–253, September 2018. ISSN 2211-3355. doi: 10.1016/j.pmedr.2018.07.005. URL <https://www.sciencedirect.com/science/article/pii/S2211335518301189>.
- [386] Tanja Heuer, Ina Schiering, and Reinhard Gerndt. Privacy-centered design for social robots. *Interaction Studies*, 20:509–529, November 2019. doi: 10.1075/is.18063.heu.
- [387] Anna Chatzimichali, Ross Harrison, and Dimitrios Chrysostomou. Toward privacy-sensitive human–robot interaction: Privacy terms and human–data interaction in the personal robot era. *Paladyn, Journal of Behavioral Robotics*, 12(1):160–174, January 2021. ISSN 2081-4836. doi: 10.1515/pjbr-2021-0013. URL <http://www.degruyter.com/document/doi/10.1515/pjbr-2021-0013/html?lang=en>. Publisher: De Gruyter Open Access.
- [388] Matthew Rueben, Alexander Mois Aroyo, Christoph Lutz, Johannes Schmölz, Pieter Van Cleynenbreugel, Andrea Corti, Siddharth Agrawal, and William D. Smart. Themes and Research Directions in Privacy-Sensitive Robotics. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 77–84, September 2018. doi: 10.1109/ARSO.2018.8625758. ISSN: 2162-7576.
- [389] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00530. arXiv: 1711.09001 ISBN: 9781538664209.

CURRICULUM VITÆ

José David VARGAS QUIRÓS

22/08/1993 Born in Gothenburg, Sweden







Education

- 2018 - 2024 **PhD Computer Science**
Delft University of Technology
Promotors: dr. H Hung and Prof. dr. ir. M.J.T. Reinders
- 2016 - 2018 **Masters in Computer Science**
Utrecht University
Supervisor: Prof. dr. A.P.J.M. Siebes
- 2011 - 2014 **Bachelor of Engineering, Electrical Engineering**
Universidad de Costa Rica

Work Experience

- 2023 - present **Erasmus Medical Center** Rotterdam, The Netherlands
Postdoc, Department of Ophtalmology
- 2023 - 2024 **Delft University of Technology** Delft, The Netherlands
Research Software Engineer, Department of Intelligent Systems
- 2018 - 2022 **Delft University of Technology** Delft, The Netherlands
PhD Candidate, Socially Perceptive Computing Lab, Department
of Intelligent Systems, EEMCS
- 2015 - 2016 **Hewlett Packard Enterprise** Heredia, Costa Rica
R&D Engineer
- 2011 - 2014 **Editec Web Development** Turrialba, Costa Rica
Freelance web developer

LIST OF PUBLICATIONS

1. Kapcak, O., **Vargas Quiros, J.D.** & Hung, H. (2019). *Estimating Romantic, Social, and Sexual Attraction by Quantifying Bodily Coordination using Wearable Sensors*. Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 154-160. <https://doi.org/10.1109/ACIIW.2019.8925137>
2. **Vargas Quiros, J.D.** & Hung, H. (2019). *CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection* Proceedings of the MediaEval 2019 Multimedia Benchmark Workshop https://ceur-ws.org/Vol-2670/MediaEval_19_paper_23.pdf
3. Cabrera-Quiros, L.C., **Vargas Quiros, J.D.**, & Hung, H. (2019). *No-Audio Multimodal Speech Detection Task at MediaEval 2020* Proceedings of the MediaEval 2020 Multimedia Benchmark Workshop <https://ceur-ws.org/Vol-2882/paper3.pdf>
4.  **Vargas Quiros, J.D.**, Kapcak, O., Hung, H. & Cabrera-Quiros, L. (2023). *Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates*. IEEE Transactions on Affective Computing, 14(3), 2168-2181. <https://doi.org/10.1109/TAFFC.2021.3138349>
5.  **Vargas-Quiros, J.V.**, Cabrera-Quiros, L., & Hung, H. (2022). *No-audio speaking status detection in crowded settings via visual pose-based filtering and wearable acceleration*. arXiv preprint: <https://arxiv.org/abs/2211.00549>
6.  **Vargas Quiros, J.D.**, Tan, S., Raman, C., Cabrera-Quiros, L. & Hung, H. (2022). *Covfee: an extensible web framework for continuous-time annotation of human behavior*. Understanding Social Behavior in Dyadic and Small Group Interactions, in Proceedings of Machine Learning Research 173:265-293. <https://proceedings.mlr.press/v173/vargas-quiros22a.html>.
7.  Raman, C*, **Vargas Quiros, J.D.***, Tan, S*, Islam, A., Gedik, E. & Hung, H. (2022). *ConfLab: a data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild*. Advances in Neural Information Processing Systems, 35, 23701-23715. <https://arxiv.org/abs/2205.05177>
8.  **Vargas-Quiros, J.D.**, Cabrera-Quiros, L., Oertel, C. & Hung, H. (2023). *Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild*. IEEE Transactions on Affective Computing, vol. 15, no. 2, pp. 519-534. <https://doi.org/10.1109/TAFFC.2023.3269003>
9.  **Vargas Quiros, J.D.**, Raman, C., Tan, S., Gedik, E., Cabrera-Quiros, L., & Hung, H. (2024). *REWIND Dataset: Privacy-preserving Speaking Status Segmentation from Multimodal Body Movement Signals in the Wild*. arXiv preprint: <https://arxiv.org/abs/2403.01229>

